Vilnius University

Remigijus Lapinskas

# A Very Short Introduction to Statistics with GRETL

remisorama@gmail.com
http://uosis.mif.vu.lt/~rlapinskas

Vilnius
2014.02

## Contents

# 1. Descriptive Statistics

These notes are currently in http://uosis.mif.vu.lt/~rlapinskas/, they are accompanied by the data sets placed in http://uosis.mif.vu.lt/~rlapinskas/ShortGRETLdata/. Prior to starting working with GRETL, create a new folder ShortINTRO on the desktop of your machine and import the two above mentioned objects into it; to place your work results, add a new folder ShortGRETLwork inside and download GRETL from http://gretl.sourceforge.net/.

****************************************

The basic idea of statistics is simple: you want to extrapolate from the data you have collected to make general conclusions about the larger population from which the data sample was derived. To do this, statisticians have developed methods based on a simple model: assume that all your data are randomly sampled from an infinitely large population. Analyze this sample, and use the results to make inferences about the population.

In statistics, we usually deal with very big data files which cannot be analysed manually. For example, below you can see 709 observations of mother‘s and father‘s height and weight (the data are contained in the file .../data/parents.txt). Is it true that men are, in general, taller than women? How should we understand the words „in general“? How can you use the data to substantiate your answer? Does the weight depend on height? Does weight depend on smoking? And what does it mean „depend“? etc. In what follows, we shall answer to some of these questions.

Here is the dataset parents.txt where

| | |
|---|---|
| `ht` | mother‘s height (in inches) |
| `dht` | dad‘s height |
| `wt` | mother‘s weight (in pounds) |
| `dwt` | dad‘s weight |
| `smoke` | 1 – mother smokes, 0 – does not smoke |

| id | wt | ht | dwt | dht | smoke |
|---|---|---|---|---|---|
| 1 | 100 | 62 | 110 | 65 | 0 |
| 2 | 135 | 64 | 148 | 70 | 0 |
| 3 | 190 | 69 | 197 | 68 | 1 |
| 4 | 93 | 62 | 130 | 64 | 1 |
| 5 | 140 | 65 | 192 | 71 | 0 |
| 6 | 125 | 62 | 180 | 70 | 0 |
| 7 | 124 | 64 | 185 | 74 | 1 |
| 8 | 130 | 63 | 205 | 71 | 0 |
| 9 | 125 | 60 | 140 | 70 | 1 |
| 10 | 142 | 66 | 195 | 73 | 1 |
| 11 | 175 | 67 | 180 | 73 | 1 |
| 12 | 145 | 66 | 150 | 70 | 1 |
| 13 | 182 | 68 | 196 | 73 | 0 |
| 14 | 106 | 58 | 200 | 68 | 1 |
| 15 | 125 | 65 | 135 | 67 | 0 |
| 16 | 132 | 66 | 168 | 70 | 1 |
| 17 | 146 | 61 | 140 | 70 | 1 |
| 18 | 123 | 66 | 210 | 66 | 1 |
| 19 | 105 | 60 | 190 | 70 | 0 |
| 20 | 130 | 67 | 185 | 71 | 0 |
| 21 | 115 | 63 | 160 | 71 | 1 |
| 22 | 92 | 63 | 178 | 71 | 1 |
| 23 | 101 | 65 | 200 | 71 | 1 |
| 24 | 160 | 61 | 130 | 67 | 0 |
| 25 | 119 | 61 | 178 | 66 | 1 |
| 26 | 130 | 65 | 155 | 72 | 1 |
| 27 | 150 | 66 | 150 | 69 | 1 |
| 28 | 90 | 60 | 150 | 66 | 0 |
| 29 | 147 | 66 | 170 | 71 | 0 |
| 30 | 119 | 63 | 165 | 67 | 1 |
| 31 | 148 | 65 | 165 | 72 | 1 |
| 32 | 126 | 64 | 200 | 69 | 0 |
| 33 | 132 | 67 | 160 | 71 | 1 |
| 34 | 130 | 60 | 165 | 70 | 0 |
| 35 | 145 | 70 | 190 | 73 | 1 |
| 36 | 140 | 65 | 195 | 69 | 1 |
| 37 | 116 | 60 | 189 | 72 | 0 |
| 38 | 96 | 61 | 170 | 71 | 0 |
| 39 | 118 | 67 | 195 | 76 | 0 |
| 40 | 130 | 63 | 174 | 72 | 1 |
| 41 | 125 | 63 | 180 | 71 | 1 |
| 42 | 115 | 65 | 192 | 68 | 1 |
| 43 | 150 | 63 | 168 | 74 | 0 |
| 44 | 137 | 69 | 170 | 72 | 1 |
| 45 | 170 | 63 | 170 | 71 | 1 |
| 46 | 170 | 63 | 165 | 69 | 1 |
| 47 | 118 | 63 | 190 | 69 | 0 |
| 48 | 125 | 66 | 165 | 70 | 1 |
| 49 | 120 | 62 | 175 | 72 | 0 |
| 50 | 110 | 62 | 163 | 69 | 0 |
| 51 | 107 | 63 | 160 | 65 | 1 |
| 52 | 130 | 63 | 155 | 73 | 0 |
| 53 | 103 | 62 | 198 | 72 | 1 |
| 54 | 116 | 65 | 168 | 66 | 1 |
| 55 | 104 | 64 | 180 | 70 | 0 |
| 56 | 135 | 68 | 160 | 73 | 1 |
| 57 | 113 | 67 | 170 | 74 | 1 |
| 58 | 112 | 62 | 150 | 66 | 0 |
| 59 | 103 | 61 | 145 | 71 | 1 |
| 60 | 100 | 64 | 210 | 71 | 0 |
| 61 | 162 | 62 | 165 | 74 | 0 |
| 62 | 110 | 62 | 168 | 71 | 1 |
| 63 | 137 | 64 | 185 | 70 | 1 |
| 64 | 120 | 64 | 155 | 74 | 0 |
| 65 | 143 | 66 | 180 | 70 | 0 |
| 66 | 125 | 65 | 160 | 70 | 0 |
| 67 | 145 | 66 | 157 | 71 | 1 |
| 68 | 114 | 64 | 150 | 70 | 1 |
| 69 | 215 | 67 | 167 | 73 | 0 |
| 70 | 145 | 66 | 185 | 72 | 1 |
| 71 | 170 | 65 | 235 | 69 | 1 |
| 72 | 133 | 66 | 170 | 73 | 0 |
| 73 | 130 | 67 | 175 | 70 | 0 |
| 74 | 155 | 69 | 170 | 71 | 0 |
| 75 | 150 | 69 | 150 | 74 | 0 |
| 76 | 150 | 61 | 150 | 67 | 1 |
| 77 | 120 | 62 | 186 | 67 | 1 |
| 78 | 154 | 66 | 175 | 69 | 0 |
| 79 | 103 | 62 | 158 | 68 | 0 |
| 80 | 100 | 60 | 153 | 71 | 0 |
| 81 | 107 | 62 | 128 | 68 | 1 |
| 82 | 123 | 65 | 180 | 70 | 1 |
| 83 | 127 | 63 | 160 | 66 | 0 |
| 84 | 155 | 66 | 165 | 71 | 0 |
| 85 | 125 | 63 | 165 | 72 | 0 |
| 86 | 175 | 71 | 205 | 74 | 1 |
| 87 | 140 | 68 | 185 | 73 | 1 |
| 88 | 250 | 66 | 181 | 69 | 0 |

| | | | | | |
|---|---|---|---|---|---|
| 89 | 148 | 68 | 165 | 72 | 0 |
| 90 | 132 | 66 | 170 | 70 | 0 |
| 91 | 152 | 64 | 155 | 64 | 0 |
| 92 | 121 | 60 | 180 | 67 | 0 |
| 93 | 138 | 62 | 185 | 71 | 0 |
| 94 | 123 | 63 | 167 | 71 | 0 |
| 95 | 160 | 65 | 156 | 67 | 1 |
| 96 | 123 | 69 | 145 | 72 | 1 |
| 97 | 123 | 65 | 160 | 70 | 1 |
| 98 | 109 | 62 | 165 | 70 | 0 |
| 99 | 115 | 65 | 165 | 73 | 0 |
| 100 | 105 | 61 | 138 | 66 | 0 |
| 101 | 131 | 66 | 172 | 72 | 0 |
| 102 | 155 | 72 | 220 | 72 | 1 |
| 103 | 170 | 66 | 225 | 72 | 0 |
| 104 | 125 | 62 | 149 | 69 | 0 |
| 105 | 120 | 64 | 165 | 69 | 1 |
| 106 | 116 | 64 | 185 | 67 | 0 |
| 107 | 220 | 63 | 220 | 74 | 0 |
| 108 | 117 | 63 | 175 | 72 | 1 |
| 109 | 93 | 61 | 165 | 68 | 0 |
| 110 | 97 | 58 | 192 | 74 | 0 |
| 111 | 135 | 59 | 165 | 68 | 1 |
| 112 | 110 | 63 | 195 | 72 | 1 |
| 113 | 124 | 63 | 157 | 73 | 1 |
| 114 | 155 | 65 | 164 | 72 | 1 |
| 115 | 150 | 63 | 140 | 66 | 0 |
| 116 | 168 | 63 | 156 | 66 | 1 |
| 117 | 147 | 66 | 208 | 73 | 1 |
| 118 | 110 | 61 | 180 | 60 | 0 |
| 119 | 140 | 63 | 150 | 66 | 0 |
| 120 | 132 | 64 | 185 | 73 | 0 |
| 121 | 105 | 61 | 140 | 69 | 0 |
| 122 | 150 | 68 | 192 | 72 | 1 |
| 123 | 125 | 65 | 150 | 72 | 1 |
| 124 | 150 | 63 | 145 | 68 | 1 |
| 125 | 138 | 62 | 235 | 76 | 1 |
| 126 | 115 | 64 | 172 | 71 | 0 |
| 127 | 125 | 65 | 160 | 72 | 0 |
| 128 | 145 | 70 | 230 | 70 | 1 |
| 129 | 130 | 65 | 175 | 71 | 1 |
| 130 | 135 | 62 | 160 | 68 | 1 |
| 131 | 121 | 63 | 145 | 68 | 1 |
| 132 | 145 | 64 | 177 | 66 | 1 |
| 133 | 136 | 68 | 145 | 69 | 0 |
| 134 | 106 | 62 | 135 | 67 | 1 |
| 135 | 117 | 66 | 212 | 74 | 0 |
| 136 | 127 | 64 | 170 | 73 | 1 |
| 137 | 160 | 63 | 200 | 68 | 0 |
| 138 | 120 | 63 | 160 | 69 | 0 |
| 139 | 110 | 63 | 145 | 66 | 0 |
| 140 | 190 | 65 | 165 | 68 | 1 |
| 141 | 140 | 65 | 180 | 71 | 1 |
| 142 | 125 | 65 | 212 | 74 | 1 |
| 143 | 117 | 67 | 212 | 76 | 1 |
| 144 | 125 | 62 | 170 | 71 | 0 |
| 145 | 124 | 65 | 115 | 62 | 1 |
| 146 | 169 | 68 | 200 | 74 | 1 |
| 147 | 125 | 63 | 165 | 72 | 0 |
| 148 | 118 | 64 | 157 | 71 | 1 |
| 149 | 139 | 64 | 150 | 66 | 0 |
| 150 | 160 | 66 | 160 | 69 | 0 |
| 151 | 135 | 66 | 155 | 68 | 0 |
| 152 | 148 | 64 | 190 | 73 | 1 |
| 153 | 122 | 65 | 180 | 70 | 0 |
| 154 | 140 | 66 | 157 | 71 | 1 |
| 155 | 132 | 66 | 168 | 71 | 0 |
| 156 | 140 | 67 | 162 | 71 | 1 |
| 157 | 103 | 61 | 150 | 72 | 0 |
| 158 | 165 | 66 | 190 | 71 | 1 |
| 159 | 115 | 62 | 145 | 67 | 1 |
| 160 | 160 | 65 | 190 | 72 | 1 |
| 161 | 140 | 59 | 140 | 68 | 1 |
| 162 | 146 | 68 | 194 | 72 | 1 |
| 163 | 108 | 63 | 173 | 72 | 0 |
| 164 | 122 | 64 | 195 | 69 | 1 |
| 165 | 152 | 71 | 173 | 71 | 1 |
| 166 | 100 | 60 | 145 | 68 | 1 |
| 167 | 125 | 62 | 195 | 71 | 0 |
| 168 | 125 | 64 | 197 | 75 | 1 |
| 169 | 102 | 61 | 156 | 68 | 1 |
| 170 | 135 | 66 | 160 | 65 | 1 |
| 171 | 145 | 62 | 149 | 66 | 1 |
| 172 | 176 | 68 | 197 | 73 | 1 |
| 173 | 130 | 66 | 150 | 67 | 1 |
| 174 | 100 | 62 | 128 | 69 | 1 |
| 175 | 110 | 63 | 146 | 68 | 1 |
| 176 | 129 | 66 | 175 | 69 | 1 |
| 177 | 136 | 66 | 217 | 73 | 0 |
| 178 | 145 | 64 | 165 | 70 | 0 |
| 179 | 120 | 66 | 165 | 74 | 0 |
| 180 | 120 | 63 | 155 | 70 | 0 |
| 181 | 120 | 60 | 140 | 61 | 1 |
| 182 | 135 | 67 | 165 | 72 | 0 |
| 183 | 132 | 63 | 170 | 69 | 0 |
| 184 | 135 | 67 | 185 | 75 | 1 |
| 185 | 127 | 64 | 165 | 67 | 0 |
| 186 | 103 | 59 | 160 | 66 | 0 |
| 187 | 157 | 63 | 170 | 71 | 1 |
| 188 | 144 | 67 | 175 | 74 | 1 |
| 189 | 130 | 68 | 160 | 71 | 1 |
| 190 | 130 | 61 | 205 | 72 | 1 |
| 191 | 130 | 67 | 195 | 72 | 0 |
| 192 | 130 | 65 | 180 | 73 | 0 |
| 193 | 103 | 63 | 170 | 68 | 1 |
| 194 | 110 | 64 | 220 | 73 | 1 |
| 195 | 122 | 59 | 152 | 69 | 0 |
| 196 | 128 | 63 | 195 | 71 | 1 |
| 197 | 132 | 67 | 205 | 74 | 1 |
| 198 | 104 | 64 | 205 | 73 | 0 |
| 199 | 115 | 64 | 193 | 73 | 0 |
| 200 | 115 | 62 | 175 | 68 | 1 |
| 201 | 110 | 64 | 165 | 69 | 1 |
| 202 | 130 | 66 | 162 | 69 | 0 |
| 203 | 130 | 66 | 200 | 70 | 0 |
| 204 | 170 | 62 | 179 | 72 | 1 |
| 205 | 122 | 62 | 170 | 74 | 0 |
| 206 | 122 | 62 | 158 | 74 | 1 |
| 207 | 108 | 60 | 160 | 63 | 0 |
| 208 | 105 | 62 | 140 | 67 | 0 |
| 209 | 125 | 65 | 165 | 71 | 0 |
| 210 | 100 | 67 | 190 | 74 | 0 |
| 211 | 137 | 67 | 200 | 75 | 0 |
| 212 | 115 | 66 | 190 | 72 | 1 |
| 213 | 112 | 65 | 170 | 71 | 1 |
| 214 | 130 | 66 | 175 | 71 | 0 |
| 215 | 145 | 68 | 190 | 74 | 0 |
| 216 | 100 | 59 | 187 | 74 | 1 |
| 217 | 140 | 64 | 200 | 73 | 1 |
| 218 | 105 | 62 | 173 | 71 | 0 |
| 219 | 104 | 60 | 145 | 64 | 0 |
| 220 | 105 | 62 | 170 | 71 | 0 |
| 221 | 120 | 62 | 147 | 65 | 0 |
| 222 | 139 | 69 | 150 | 69 | 0 |
| 223 | 116 | 64 | 180 | 75 | 1 |
| 224 | 124 | 66 | 178 | 74 | 0 |
| 225 | 143 | 64 | 175 | 73 | 0 |
| 226 | 137 | 61 | 160 | 74 | 1 |
| 227 | 132 | 67 | 215 | 75 | 0 |
| 228 | 130 | 67 | 205 | 72 | 0 |
| 229 | 155 | 66 | 173 | 69 | 1 |
| 230 | 138 | 63 | 170 | 67 | 1 |
| 231 | 102 | 62 | 165 | 72 | 1 |
| 232 | 140 | 68 | 170 | 71 | 0 |
| 233 | 148 | 66 | 165 | 66 | 0 |
| 234 | 135 | 66 | 155 | 74 | 0 |
| 235 | 120 | 61 | 190 | 72 | 1 |
| 236 | 130 | 63 | 145 | 67 | 0 |
| 237 | 115 | 65 | 165 | 70 | 0 |
| 238 | 124 | 63 | 155 | 68 | 1 |
| 239 | 118 | 62 | 190 | 73 | 1 |
| 240 | 137 | 67 | 165 | 70 | 0 |
| 241 | 118 | 66 | 205 | 71 | 1 |
| 242 | 150 | 61 | 170 | 70 | 0 |
| 243 | 127 | 65 | 183 | 69 | 1 |
| 244 | 138 | 65 | 190 | 70 | 1 |
| 245 | 93 | 60 | 180 | 75 | 1 |
| 246 | 125 | 61 | 175 | 68 | 1 |
| 247 | 123 | 65 | 150 | 71 | 0 |
| 248 | 111 | 66 | 210 | 72 | 1 |
| 249 | 133 | 67 | 180 | 74 | 0 |
| 250 | 147 | 64 | 240 | 73 | 0 |
| 251 | 125 | 64 | 185 | 73 | 0 |
| 252 | 115 | 60 | 150 | 68 | 1 |
| 253 | 135 | 66 | 178 | 72 | 1 |
| 254 | 110 | 63 | 200 | 76 | 1 |
| 255 | 130 | 64 | 140 | 70 | 1 |
| 256 | 135 | 63 | 190 | 72 | 0 |
| 257 | 128 | 63 | 180 | 71 | 1 |
| 258 | 118 | 65 | 150 | 69 | 0 |
| 259 | 123 | 63 | 195 | 71 | 0 |
| 260 | 125 | 64 | 155 | 70 | 0 |
| 261 | 120 | 64 | 145 | 69 | 1 |
| 262 | 122 | 67 | 163 | 68 | 1 |
| 263 | 103 | 59 | 142 | 64 | 1 |
| 264 | 110 | 63 | 190 | 72 | 1 |
| 265 | 130 | 66 | 183 | 73 | 1 |
| 266 | 117 | 64 | 168 | 72 | 1 |
| 267 | 95 | 62 | 150 | 67 | 0 |
| 268 | 126 | 67 | 190 | 71 | 0 |
| 269 | 147 | 65 | 145 | 68 | 0 |
| 270 | 140 | 63 | 150 | 66 | 0 |
| 271 | 180 | 64 | 180 | 72 | 1 |
| 272 | 102 | 59 | 163 | 66 | 0 |
| 273 | 116 | 64 | 195 | 75 | 1 |
| 274 | 110 | 62 | 120 | 66 | 1 |
| 275 | 115 | 65 | 150 | 72 | 1 |
| 276 | 145 | 66 | 215 | 74 | 0 |
| 277 | 140 | 65 | 165 | 69 | 1 |
| 278 | 125 | 66 | 205 | 72 | 1 |
| 279 | 180 | 66 | 140 | 68 | 1 |
| 280 | 120 | 64 | 180 | 67 | 0 |
| 281 | 109 | 60 | 250 | 69 | 1 |
| 282 | 113 | 64 | 197 | 72 | 1 |
| 283 | 132 | 63 | 180 | 74 | 0 |
| 284 | 110 | 62 | 145 | 68 | 0 |
| 285 | 160 | 65 | 165 | 73 | 0 |
| 286 | 103 | 62 | 175 | 70 | 1 |
| 287 | 128 | 66 | 162 | 72 | 0 |
| 288 | 96 | 59 | 170 | 71 | 0 |
| 289 | 120 | 66 | 170 | 71 | 1 |
| 290 | 108 | 64 | 150 | 65 | 1 |
| 291 | 145 | 66 | 180 | 70 | 0 |
| 292 | 140 | 59 | 190 | 72 | 1 |
| 293 | 135 | 66 | 192 | 70 | 0 |
| 294 | 155 | 67 | 175 | 72 | 0 |
| 295 | 105 | 65 | 156 | 69 | 1 |
| 296 | 102 | 63 | 135 | 67 | 0 |
| 297 | 124 | 65 | 170 | 71 | 1 |
| 298 | 145 | 66 | 165 | 67 | 0 |
| 299 | 130 | 61 | 175 | 71 | 0 |
| 300 | 135 | 65 | 188 | 72 | 1 |
| 301 | 134 | 68 | 190 | 74 | 1 |
| 302 | 105 | 64 | 220 | 73 | 1 |
| 303 | 132 | 65 | 180 | 67 | 0 |
| 304 | 150 | 65 | 185 | 71 | 0 |
| 305 | 105 | 62 | 170 | 68 | 1 |
| 306 | 113 | 60 | 160 | 66 | 0 |
| 307 | 98 | 59 | 146 | 69 | 1 |
| 308 | 135 | 65 | 155 | 68 | 1 |
| 309 | 130 | 67 | 148 | 70 | 1 |
| 310 | 120 | 63 | 155 | 69 | 0 |
| 311 | 129 | 64 | 180 | 71 | 1 |
| 312 | 115 | 66 | 164 | 67 | 0 |
| 313 | 107 | 63 | 190 | 72 | 0 |
| 314 | 115 | 62 | 185 | 69 | 1 |
| 315 | 110 | 64 | 165 | 70 | 1 |
| 316 | 137 | 64 | 164 | 71 | 1 |
| 317 | 115 | 64 | 170 | 70 | 1 |
| 318 | 139 | 68 | 185 | 75 | 1 |
| 319 | 140 | 65 | 160 | 68 | 1 |
| 320 | 100 | 61 | 130 | 72 | 0 |
| 321 | 160 | 69 | 202 | 73 | 1 |
| 322 | 108 | 62 | 185 | 68 | 0 |
| 323 | 132 | 62 | 147 | 67 | 0 |
| 324 | 165 | 69 | 200 | 70 | 0 |
| 325 | 109 | 62 | 135 | 69 | 1 |
| 326 | 110 | 62 | 165 | 64 | 1 |
| 327 | 202 | 63 | 170 | 69 | 1 |
| 328 | 112 | 58 | 156 | 68 | 1 |
| 329 | 108 | 62 | 180 | 71 | 1 |
| 330 | 125 | 61 | 220 | 76 | 0 |
| 331 | 180 | 68 | 200 | 73 | 0 |
| 332 | 130 | 68 | 156 | 72 | 1 |
| 333 | 90 | 59 | 148 | 72 | 1 |
| 334 | 118 | 62 | 185 | 72 | 1 |
| 335 | 120 | 61 | 180 | 68 | 1 |
| 336 | 145 | 65 | 215 | 75 | 1 |
| 337 | 129 | 66 | 165 | 72 | 0 |
| 338 | 112 | 64 | 170 | 71 | 1 |
| 339 | 155 | 64 | 175 | 71 | 0 |
| 340 | 124 | 67 | 182 | 71 | 1 |
| 341 | 132 | 68 | 180 | 70 | 0 |
| 342 | 112 | 63 | 170 | 72 | 0 |
| 343 | 101 | 65 | 146 | 65 | 1 |
| 344 | 117 | 65 | 165 | 70 | 1 |
| 345 | 134 | 67 | 180 | 71 | 0 |
| 346 | 125 | 64 | 193 | 72 | 0 |
| 347 | 108 | 63 | 180 | 76 | 1 |
| 348 | 145 | 68 | 180 | 70 | 0 |
| 349 | 132 | 65 | 180 | 73 | 1 |
| 350 | 115 | 60 | 150 | 69 | 1 |
| 351 | 118 | 64 | 190 | 71 | 1 |
| 352 | 116 | 64 | 125 | 68 | 0 |

| | | | | | |
|---|---|---|---|---|---|
| 353 | 145 | 70 | 165 | 70 | 1 |
| 354 | 125 | 61 | 160 | 70 | 0 |
| 355 | 115 | 60 | 175 | 66 | 1 |
| 356 | 120 | 65 | 148 | 70 | 1 |
| 357 | 94 | 62 | 159 | 67 | 0 |
| 358 | 105 | 61 | 158 | 68 | 0 |
| 359 | 120 | 62 | 182 | 73 | 1 |
| 360 | 117 | 60 | 150 | 67 | 1 |
| 361 | 126 | 65 | 145 | 72 | 1 |
| 362 | 175 | 67 | 148 | 61 | 0 |
| 363 | 169 | 68 | 195 | 76 | 0 |
| 364 | 135 | 66 | 205 | 72 | 1 |
| 365 | 120 | 64 | 155 | 72 | 0 |
| 366 | 125 | 61 | 162 | 66 | 0 |
| 367 | 128 | 66 | 170 | 71 | 1 |
| 368 | 118 | 66 | 170 | 68 | 1 |
| 369 | 150 | 58 | 160 | 67 | 1 |
| 370 | 115 | 62 | 165 | 68 | 1 |
| 371 | 122 | 64 | 160 | 69 | 0 |
| 372 | 127 | 64 | 200 | 72 | 0 |
| 373 | 160 | 64 | 200 | 69 | 1 |
| 374 | 130 | 62 | 190 | 73 | 0 |
| 375 | 198 | 68 | 145 | 69 | 0 |
| 376 | 190 | 63 | 215 | 71 | 0 |
| 377 | 130 | 69 | 180 | 73 | 1 |
| 378 | 100 | 62 | 140 | 71 | 1 |
| 379 | 114 | 67 | 178 | 73 | 0 |
| 380 | 136 | 69 | 175 | 74 | 1 |
| 381 | 120 | 64 | 160 | 66 | 0 |
| 382 | 96 | 59 | 168 | 69 | 0 |
| 383 | 110 | 62 | 188 | 72 | 1 |
| 384 | 182 | 61 | 150 | 72 | 0 |
| 385 | 122 | 66 | 195 | 72 | 0 |
| 386 | 135 | 64 | 175 | 72 | 1 |
| 387 | 125 | 61 | 173 | 69 | 0 |
| 388 | 127 | 64 | 180 | 71 | 0 |
| 389 | 125 | 66 | 175 | 75 | 1 |
| 390 | 120 | 62 | 150 | 70 | 1 |
| 391 | 118 | 62 | 190 | 69 | 0 |
| 392 | 147 | 68 | 175 | 70 | 1 |
| 393 | 127 | 64 | 182 | 76 | 1 |
| 394 | 134 | 62 | 138 | 64 | 0 |
| 395 | 121 | 62 | 150 | 68 | 1 |
| 396 | 120 | 63 | 165 | 68 | 0 |
| 397 | 110 | 62 | 180 | 74 | 0 |
| 398 | 160 | 68 | 135 | 69 | 1 |
| 399 | 145 | 63 | 165 | 72 | 1 |
| 400 | 96 | 60 | 145 | 65 | 0 |
| 401 | 130 | 63 | 140 | 69 | 0 |
| 402 | 119 | 64 | 175 | 70 | 1 |
| 403 | 124 | 66 | 156 | 72 | 0 |
| 404 | 130 | 67 | 180 | 77 | 1 |
| 405 | 164 | 68 | 183 | 71 | 0 |
| 406 | 155 | 65 | 170 | 75 | 0 |
| 407 | 149 | 64 | 142 | 71 | 1 |
| 408 | 139 | 66 | 160 | 71 | 1 |
| 409 | 130 | 68 | 195 | 74 | 0 |
| 410 | 110 | 63 | 185 | 69 | 0 |
| 411 | 140 | 64 | 190 | 71 | 1 |
| 412 | 124 | 65 | 160 | 69 | 1 |
| 413 | 129 | 67 | 180 | 73 | 1 |
| 414 | 132 | 63 | 135 | 71 | 1 |
| 415 | 132 | 69 | 180 | 72 | 0 |
| 416 | 145 | 66 | 125 | 67 | 0 |
| 417 | 124 | 64 | 165 | 72 | 0 |
| 418 | 130 | 65 | 165 | 72 | 1 |
| 419 | 110 | 65 | 170 | 67 | 1 |
| 420 | 108 | 61 | 127 | 64 | 1 |
| 421 | 162 | 66 | 162 | 72 | 0 |
| 422 | 130 | 66 | 183 | 72 | 0 |
| 423 | 118 | 65 | 155 | 67 | 0 |
| 424 | 103 | 63 | 160 | 71 | 1 |
| 425 | 107 | 61 | 167 | 73 | 1 |
| 426 | 112 | 63 | 187 | 70 | 1 |
| 427 | 141 | 67 | 170 | 70 | 1 |
| 428 | 118 | 63 | 145 | 70 | 1 |
| 429 | 148 | 70 | 205 | 75 | 1 |
| 430 | 125 | 64 | 220 | 75 | 0 |
| 431 | 120 | 67 | 185 | 73 | 0 |
| 432 | 122 | 61 | 190 | 71 | 0 |
| 433 | 126 | 64 | 208 | 67 | 0 |
| 434 | 108 | 61 | 154 | 68 | 1 |
| 435 | 107 | 60 | 135 | 69 | 0 |
| 436 | 215 | 67 | 170 | 72 | 0 |
| 437 | 110 | 64 | 157 | 67 | 1 |
| 438 | 150 | 64 | 153 | 71 | 1 |
| 439 | 130 | 65 | 160 | 71 | 0 |
| 440 | 107 | 64 | 150 | 67 | 0 |
| 441 | 191 | 66 | 148 | 66 | 0 |
| 442 | 112 | 62 | 150 | 70 | 0 |
| 443 | 110 | 66 | 193 | 74 | 1 |
| 444 | 140 | 66 | 155 | 68 | 1 |
| 445 | 210 | 66 | 180 | 71 | 1 |
| 446 | 160 | 64 | 205 | 72 | 0 |
| 447 | 143 | 66 | 160 | 70 | 0 |
| 448 | 125 | 61 | 145 | 68 | 0 |
| 449 | 100 | 62 | 180 | 69 | 0 |
| 450 | 129 | 63 | 155 | 71 | 1 |
| 451 | 150 | 64 | 210 | 69 | 1 |
| 452 | 110 | 61 | 156 | 67 | 0 |
| 453 | 125 | 65 | 130 | 72 | 1 |
| 454 | 118 | 63 | 158 | 70 | 1 |
| 455 | 181 | 68 | 180 | 69 | 0 |
| 456 | 117 | 66 | 187 | 71 | 0 |
| 457 | 135 | 67 | 195 | 74 | 1 |
| 458 | 130 | 66 | 188 | 73 | 1 |
| 459 | 110 | 66 | 160 | 68 | 1 |
| 460 | 110 | 67 | 165 | 70 | 0 |
| 461 | 140 | 67 | 160 | 66 | 1 |
| 462 | 120 | 60 | 120 | 65 | 1 |
| 463 | 155 | 67 | 175 | 72 | 0 |
| 464 | 112 | 64 | 160 | 71 | 0 |
| 465 | 128 | 61 | 212 | 73 | 1 |
| 466 | 136 | 61 | 140 | 66 | 0 |
| 467 | 153 | 66 | 178 | 73 | 0 |
| 468 | 99 | 61 | 147 | 67 | 1 |
| 469 | 115 | 59 | 168 | 66 | 0 |
| 470 | 175 | 65 | 230 | 75 | 1 |
| 471 | 107 | 63 | 184 | 72 | 1 |
| 472 | 160 | 62 | 168 | 69 | 0 |
| 473 | 128 | 63 | 180 | 73 | 0 |
| 474 | 130 | 67 | 173 | 70 | 0 |
| 475 | 125 | 64 | 185 | 72 | 1 |
| 476 | 108 | 62 | 138 | 64 | 0 |
| 477 | 155 | 61 | 165 | 69 | 1 |
| 478 | 135 | 67 | 165 | 70 | 1 |
| 479 | 115 | 62 | 180 | 70 | 0 |
| 480 | 105 | 62 | 175 | 67 | 1 |
| 481 | 143 | 68 | 160 | 71 | 1 |
| 482 | 120 | 64 | 180 | 72 | 1 |
| 483 | 134 | 67 | 190 | 75 | 1 |
| 484 | 121 | 61 | 160 | 66 | 0 |
| 485 | 160 | 67 | 180 | 63 | 0 |
| 486 | 109 | 64 | 150 | 71 | 1 |
| 487 | 133 | 65 | 150 | 66 | 1 |
| 488 | 121 | 61 | 185 | 75 | 1 |
| 489 | 112 | 64 | 190 | 72 | 1 |
| 490 | 130 | 68 | 190 | 74 | 0 |
| 491 | 145 | 64 | 163 | 68 | 0 |
| 492 | 124 | 67 | 173 | 72 | 0 |
| 493 | 120 | 64 | 150 | 68 | 1 |
| 494 | 135 | 65 | 220 | 72 | 0 |
| 495 | 133 | 66 | 160 | 71 | 1 |
| 496 | 122 | 64 | 170 | 70 | 1 |
| 497 | 125 | 65 | 160 | 71 | 1 |
| 498 | 112 | 64 | 155 | 72 | 1 |
| 499 | 135 | 66 | 185 | 72 | 0 |
| 500 | 115 | 63 | 180 | 76 | 0 |
| 501 | 135 | 69 | 165 | 69 | 0 |
| 502 | 131 | 67 | 168 | 71 | 0 |
| 503 | 160 | 64 | 145 | 67 | 0 |
| 504 | 140 | 65 | 250 | 74 | 0 |
| 505 | 149 | 63 | 200 | 71 | 1 |
| 506 | 110 | 65 | 185 | 69 | 0 |
| 507 | 155 | 65 | 183 | 71 | 1 |
| 508 | 150 | 63 | 165 | 68 | 0 |
| 509 | 129 | 65 | 180 | 71 | 1 |
| 510 | 122 | 61 | 150 | 73 | 1 |
| 511 | 112 | 64 | 150 | 67 | 1 |
| 512 | 130 | 62 | 170 | 68 | 1 |
| 513 | 128 | 69 | 145 | 69 | 1 |
| 514 | 120 | 63 | 145 | 68 | 0 |
| 515 | 98 | 66 | 160 | 71 | 0 |
| 516 | 127 | 65 | 148 | 67 | 1 |
| 517 | 145 | 63 | 175 | 73 | 1 |
| 518 | 130 | 62 | 175 | 66 | 0 |
| 519 | 113 | 60 | 150 | 65 | 0 |
| 520 | 156 | 64 | 175 | 67 | 0 |
| 521 | 110 | 60 | 150 | 69 | 0 |
| 522 | 132 | 65 | 159 | 72 | 1 |
| 523 | 140 | 68 | 163 | 71 | 0 |
| 524 | 164 | 62 | 150 | 65 | 1 |
| 525 | 140 | 66 | 205 | 78 | 1 |
| 526 | 127 | 66 | 165 | 71 | 1 |
| 527 | 115 | 63 | 170 | 70 | 1 |
| 528 | 185 | 68 | 185 | 72 | 0 |
| 529 | 124 | 67 | 200 | 70 | 1 |
| 530 | 114 | 61 | 185 | 70 | 1 |
| 531 | 116 | 61 | 143 | 66 | 0 |
| 532 | 133 | 63 | 188 | 70 | 1 |
| 533 | 115 | 60 | 150 | 72 | 0 |
| 534 | 140 | 59 | 165 | 61 | 0 |
| 535 | 127 | 62 | 200 | 71 | 1 |
| 536 | 120 | 64 | 155 | 70 | 1 |
| 537 | 130 | 61 | 180 | 72 | 1 |
| 538 | 135 | 62 | 173 | 75 | 0 |
| 539 | 145 | 70 | 190 | 75 | 1 |
| 540 | 150 | 68 | 220 | 74 | 0 |
| 541 | 109 | 63 | 210 | 70 | 1 |
| 542 | 110 | 64 | 140 | 71 | 0 |
| 543 | 132 | 65 | 165 | 73 | 0 |
| 544 | 110 | 61 | 166 | 68 | 1 |
| 545 | 120 | 64 | 185 | 76 | 0 |
| 546 | 130 | 67 | 200 | 73 | 1 |
| 547 | 130 | 69 | 190 | 72 | 0 |
| 548 | 125 | 63 | 173 | 72 | 0 |
| 549 | 112 | 65 | 178 | 72 | 1 |
| 550 | 132 | 65 | 168 | 70 | 0 |
| 551 | 128 | 65 | 175 | 71 | 1 |
| 552 | 115 | 64 | 140 | 69 | 0 |
| 553 | 116 | 67 | 198 | 75 | 1 |
| 554 | 145 | 68 | 145 | 68 | 0 |
| 555 | 119 | 63 | 165 | 71 | 0 |
| 556 | 135 | 63 | 146 | 66 | 0 |
| 557 | 117 | 66 | 165 | 68 | 1 |
| 558 | 117 | 65 | 123 | 64 | 1 |
| 559 | 120 | 67 | 210 | 72 | 0 |
| 560 | 130 | 65 | 190 | 74 | 0 |
| 561 | 120 | 60 | 150 | 64 | 1 |
| 562 | 125 | 63 | 170 | 73 | 1 |
| 563 | 129 | 66 | 183 | 73 | 1 |
| 564 | 144 | 69 | 188 | 70 | 1 |
| 565 | 145 | 64 | 206 | 68 | 0 |
| 566 | 104 | 63 | 185 | 73 | 1 |
| 567 | 110 | 63 | 187 | 73 | 1 |
| 568 | 145 | 65 | 190 | 73 | 0 |
| 569 | 125 | 66 | 160 | 69 | 1 |
| 570 | 108 | 64 | 166 | 67 | 0 |
| 571 | 119 | 60 | 149 | 67 | 0 |
| 572 | 118 | 61 | 155 | 70 | 1 |
| 573 | 130 | 65 | 200 | 73 | 1 |
| 574 | 97 | 62 | 150 | 73 | 1 |
| 575 | 115 | 61 | 170 | 73 | 1 |
| 576 | 135 | 68 | 165 | 71 | 0 |
| 577 | 142 | 67 | 140 | 70 | 1 |
| 578 | 131 | 66 | 170 | 70 | 1 |
| 579 | 165 | 65 | 160 | 68 | 0 |
| 580 | 122 | 65 | 152 | 69 | 0 |
| 581 | 114 | 60 | 160 | 67 | 0 |
| 582 | 137 | 67 | 170 | 74 | 1 |
| 583 | 113 | 64 | 175 | 75 | 0 |
| 584 | 145 | 65 | 170 | 71 | 0 |
| 585 | 110 | 67 | 220 | 71 | 0 |
| 586 | 126 | 65 | 168 | 71 | 0 |
| 587 | 135 | 65 | 165 | 68 | 0 |
| 588 | 135 | 65 | 160 | 71 | 0 |
| 589 | 105 | 62 | 140 | 68 | 1 |
| 590 | 113 | 61 | 185 | 72 | 1 |
| 591 | 110 | 64 | 155 | 72 | 1 |
| 592 | 121 | 65 | 170 | 71 | 1 |
| 593 | 130 | 67 | 180 | 74 | 0 |
| 594 | 122 | 62 | 190 | 73 | 0 |
| 595 | 127 | 64 | 180 | 74 | 0 |
| 596 | 122 | 66 | 185 | 69 | 0 |
| 597 | 115 | 61 | 110 | 61 | 0 |
| 598 | 104 | 64 | 188 | 72 | 0 |
| 599 | 146 | 67 | 190 | 71 | 0 |
| 600 | 113 | 59 | 175 | 70 | 1 |
| 601 | 120 | 61 | 134 | 65 | 1 |
| 602 | 142 | 61 | 169 | 69 | 0 |
| 603 | 124 | 64 | 140 | 72 | 1 |
| 604 | 127 | 68 | 215 | 74 | 1 |
| 605 | 135 | 66 | 150 | 68 | 1 |
| 606 | 122 | 68 | 140 | 72 | 1 |
| 607 | 127 | 62 | 140 | 65 | 1 |
| 608 | 125 | 63 | 158 | 74 | 1 |
| 609 | 135 | 65 | 165 | 70 | 0 |
| 610 | 150 | 67 | 208 | 70 | 0 |
| 611 | 155 | 70 | 185 | 70 | 0 |
| 612 | 120 | 63 | 185 | 69 | 1 |
| 613 | 115 | 60 | 165 | 69 | 1 |
| 614 | 136 | 68 | 165 | 69 | 1 |
| 615 | 115 | 60 | 136 | 66 | 0 |
| 616 | 118 | 64 | 215 | 71 | 1 |

| | | | | | |
|---|---|---|---|---|---|
| 617 | 120 | 61 | 160 | 68 | 0 |
| 618 | 118 | 66 | 150 | 67 | 0 |
| 619 | 105 | 60 | 140 | 65 | 1 |
| 620 | 154 | 62 | 199 | 69 | 1 |
| 621 | 118 | 63 | 175 | 68 | 0 |
| 622 | 122 | 63 | 195 | 77 | 0 |
| 623 | 117 | 63 | 150 | 66 | 1 |
| 624 | 150 | 61 | 150 | 69 | 0 |
| 625 | 115 | 65 | 170 | 75 | 1 |
| 626 | 118 | 62 | 205 | 71 | 0 |
| 627 | 102 | 62 | 150 | 68 | 1 |
| 628 | 127 | 64 | 150 | 70 | 0 |
| 629 | 104 | 61 | 135 | 71 | 0 |
| 630 | 99 | 58 | 130 | 66 | 0 |
| 631 | 107 | 63 | 173 | 69 | 0 |
| 632 | 124 | 63 | 190 | 73 | 1 |
| 633 | 142 | 65 | 180 | 71 | 1 |
| 634 | 132 | 67 | 245 | 78 | 0 |
| 635 | 125 | 63 | 170 | 69 | 1 |
| 636 | 106 | 62 | 140 | 66 | 0 |
| 637 | 120 | 65 | 165 | 71 | 1 |
| 638 | 200 | 66 | 170 | 72 | 1 |
| 639 | 112 | 61 | 135 | 65 | 1 |
| 640 | 114 | 64 | 140 | 69 | 1 |
| 641 | 117 | 68 | 140 | 68 | 1 |
| 642 | 99 | 61 | 195 | 75 | 0 |
| 643 | 177 | 66 | 175 | 71 | 1 |
| 644 | 145 | 66 | 165 | 67 | 0 |
| 645 | 124 | 61 | 145 | 67 | 0 |
| 646 | 123 | 65 | 165 | 69 | 1 |
| 647 | 130 | 67 | 170 | 71 | 1 |
| 648 | 110 | 64 | 260 | 71 | 1 |
| 649 | 119 | 63 | 140 | 68 | 1 |
| 650 | 104 | 64 | 165 | 73 | 1 |
| 651 | 103 | 59 | 170 | 66 | 1 |
| 652 | 135 | 66 | 200 | 71 | 0 |
| 653 | 180 | 63 | 170 | 71 | 0 |
| 654 | 110 | 63 | 160 | 71 | 1 |
| 655 | 145 | 61 | 145 | 68 | 1 |
| 656 | 150 | 65 | 175 | 71 | 1 |
| 657 | 128 | 64 | 120 | 65 | 0 |
| 658 | 115 | 63 | 175 | 72 | 0 |
| 659 | 145 | 67 | 163 | 72 | 1 |
| 660 | 130 | 65 | 170 | 69 | 1 |
| 661 | 103 | 63 | 160 | 65 | 1 |
| 662 | 126 | 64 | 138 | 68 | 0 |
| 663 | 113 | 63 | 155 | 72 | 1 |
| 664 | 130 | 61 | 160 | 71 | 0 |
| 665 | 137 | 65 | 167 | 74 | 1 |
| 666 | 112 | 61 | 195 | 71 | 0 |
| 667 | 127 | 65 | 170 | 72 | 1 |
| 668 | 110 | 62 | 137 | 71 | 1 |
| 669 | 145 | 63 | 180 | 69 | 1 |
| 670 | 140 | 66 | 190 | 74 | 1 |
| 671 | 135 | 62 | 180 | 73 | 1 |
| 672 | 228 | 65 | 220 | 70 | 0 |
| 673 | 160 | 65 | 190 | 73 | 0 |
| 674 | 158 | 65 | 215 | 75 | 1 |
| 675 | 145 | 67 | 172 | 74 | 1 |
| 676 | 127 | 64 | 145 | 64 | 1 |
| 677 | 135 | 67 | 176 | 71 | 1 |
| 678 | 150 | 63 | 180 | 68 | 1 |
| 679 | 170 | 64 | 170 | 69 | 0 |
| 680 | 107 | 63 | 183 | 75 | 1 |
| 681 | 130 | 58 | 155 | 69 | 0 |
| 682 | 115 | 63 | 185 | 71 | 1 |
| 683 | 120 | 61 | 155 | 67 | 0 |
| 684 | 113 | 64 | 167 | 69 | 0 |
| 685 | 125 | 66 | 140 | 67 | 1 |
| 686 | 156 | 54 | 195 | 69 | 1 |
| 687 | 140 | 66 | 165 | 71 | 1 |
| 688 | 130 | 66 | 185 | 71 | 1 |
| 689 | 103 | 65 | 150 | 67 | 1 |
| 690 | 120 | 68 | 245 | 74 | 1 |
| 691 | 151 | 69 | 185 | 69 | 1 |
| 692 | 103 | 63 | 170 | 68 | 0 |
| 693 | 109 | 62 | 130 | 64 | 0 |
| 694 | 145 | 66 | 180 | 73 | 0 |
| 695 | 150 | 63 | 160 | 67 | 1 |
| 696 | 180 | 66 | 190 | 70 | 0 |
| 697 | 95 | 60 | 150 | 73 | 0 |
| 698 | 120 | 65 | 150 | 73 | 1 |
| 699 | 116 | 64 | 140 | 70 | 1 |
| 700 | 136 | 64 | 201 | 73 | 1 |
| 701 | 102 | 63 | 140 | 66 | 1 |
| 702 | 87 | 60 | 182 | 71 | 1 |
| 703 | 121 | 65 | 140 | 66 | 1 |
| 704 | 126 | 65 | 200 | 69 | 1 |
| 705 | 100 | 60 | 190 | 72 | 0 |
| 706 | 120 | 67 | 170 | 73 | 0 |
| 707 | 150 | 65 | 180 | 70 | 1 |
| 708 | 110 | 65 | 165 | 71 | 0 |
| 709 | 129 | 65 | 172 | 68 | 0 |

## 1.1. Types of Variables

Variables play different roles, and knowing the variable's *type* is crucial to knowing what to do with it and what it can tell us.

When a variable names categories and answers questions about how cases fall into those categories, we call it a **categorical**, or **qualitative, variable** (a categorical variable that names categories that don't have order is sometimes called **nominal**; `smoke` is both a categorical and nominal variable). When a variable has measured numerical values with *units* and the variable tells us about the quantity of what is measured, we call it a **quantitative variable** (height and weight are quantitative variables).

A **time series** is a single variable measured at regular intervals over time. Typical measuring points are months, quarters, or years (see Ch. 4). By contrast, most of the methods in this book are better suited for **cross-sectional data**, where several variables are measured at the same time point: if we collect data on sales revenue, number of customers, and expenses for last month at *each* Starbucks (more than 16,000 locations as of 2010) at one point in time, this would be cross-sectional data.

## 1.2. Starting GRETL

GRETL is an open-source statistical package, mainly for econometrics (econometrics is a part of statistics dealing normally with economic models and/or economic data). The name is an acronym for *G*nu *R*egression, *E*conometrics and *T*ime-series *L*ibrary. The product can be freely downloaded from [http://gretl.sourceforge.net/](http://gretl.sourceforge.net/). In this course, we shall use GRETL for introductory statistics, basic hypothesis testing and first steps in regression and time series analysis.

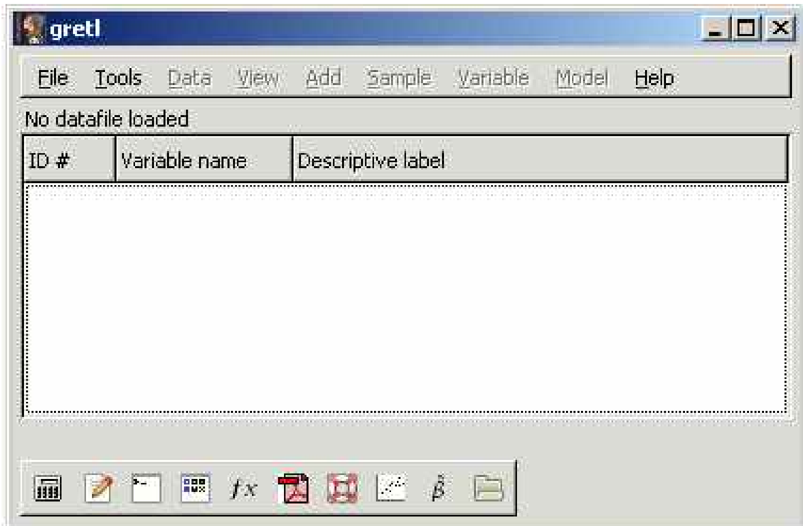



Fig. 1.1. Introductory screen

On clicking the GRETL's icon, the figure shown above appears. We assume that this is the first sesion and no data had been saved before. To import the file parents.txt, click File*Open data*Import*txt\CSV and navigate to .../ShortGRETLdata/parents.txt. Treat the data as undated.

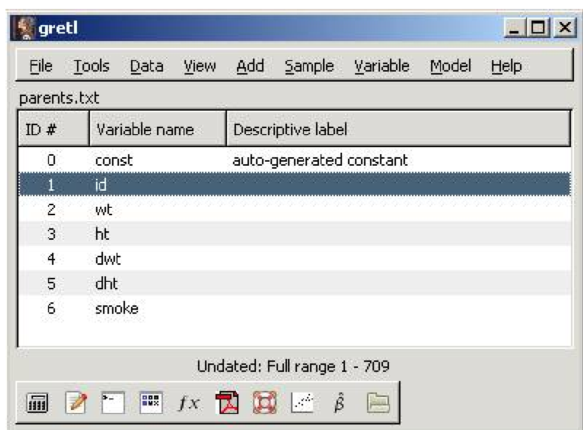



Fig. 1.2. Your data are now in GRETL

GRETL has several possibilities to perform statistical analysis.

1. Menu driven (GUI) interface.
2. Scripting (programable) approach.
3. Command line interface (commanding from console)

You can, if you wish, use the GUI controls and the scripting approach in tandem, exploiting each method where it offers greater convenience. Here are two suggestions.

- Open a data file in the GUI. Explore the data — generate graphs, run regressions, perform tests. Then open the Tools| Command log, edit out any redundant commands, and save it under a specific name. Run the script to generate a single file containing a concise record of your work.

- Start by establishing a new script file. Type in any commands that may be required to set up transformations of the data (see the genr command in the *Gretl Command Reference*). Typically this sort of thing can be accomplished more efficiently via commands assembled with forethought rather than point-and-click. Then save and run the script: the GUI data window will be updated accordingly. Now you can carry out further exploration of the data via the GUI. To revisit the data at a later point, open and rerun the "preparatory" script first.

We start with the **first option**. To look through the data, click Data*Select all, then right-click on selected variables and choose Display values – you will get the same as above, not very informative table.

```
              id          wt          ht          dwt          dht      smoke

  1            1         100          62          110           65          0
  2            2         135          64          148           70          0
  3            3         190          69          197           68          1
  4            4          93          62          130           64          1
  5            5         140          65          192           71          0
  6            6         125          62          180           70          0
  7            7         124          64          185           74          1
  8            8         130          63          205           71          0
```
..............................................................................................................................................

We would like to compact our data, to get some aggregate or **descriptive** characteristics of our data set. Let us start with

## 1.3. Numeric Characteristics

Right-click on selected variables and choose Descriptive statistics.

```
Summary statistics, using the observations 1 - 709

                         Mean         Median        Minimum        Maximum
id                     355,00         355,00         1,0000         709,00
wt                     128,87         125,00         87,000         250,00
ht                     64,080         64,000         54,000         72,000
dwt                    171,10         170,00         110,00         260,00
dht                    70,247         71,000         60,000         78,000
smoke                 0,53456         1,0000        0,00000         1,0000

                     Std. Dev.          C.V.       Skewness   Ex. kurtosis
id                     204,81        0,57694   -4,7878e-021        -1,2000
wt                     21,034        0,16321         1,3425         3,4996
ht                     2,5352       0,039563      -0,046050      -0,094970
dwt                    22,409        0,13098        0,44915        0,60613
dht                    2,8567       0,040667       -0,36147        0,16266
smoke                 0,49916        0,93378       -0,13855        -1,9808
```

In the above table, you can see some, maybe, unknown terms such as median, standard deviation and others. We start with mean and median which are used to describe the **central value** of a sample.

The **mean** is an average, one of several, that summarise the typical value of a set of data. The mean is the grand total divided by the number of data points, i.e., if our sample consists of the numbers $x_1, x_2, ..., x_n$, then its sample mean $\bar{x} = (x_1 + ... + x_n)/n$.

Note that the mean depends on the sample (if you take another sample of parents, you will get a (hopefully, only a slightly) different value of $\bar{x}$; it is termed a sampling variation). It also depends on the sample size $n$ (when $n$ increases, $\bar{x}$ becomes closer and closer to the true mean of the whole population).

The **median** is the middle value in a sample sorted into ascending order. If the sample contains an even number of values, the median is defined as the mean of the middle two.

Is it better to use the mean or the median? This may sound like an obscure technical question, but it really can matter. The short answer is "it depends" - to know which you should use, you must know how your data is distributed. The mean is the one to use with symmetrically distributed data; otherwise, use the median. If you follow this rule, you will get a more accurate[1] reflection of an "average" value.

Coming back to our data, we can already draw some conclusions: the mean of men's weight (171,10 pounds) is definitely higher that that of women (128,87). The same is true for hight: 70,247 (inches) vs 64,080 etc

Another question - is it true that smoking women weight less? Go to Tools*Test statistic calculator*2 means and fill it as shown below (in this, two subsamples case, you have to press the Enter key after printing ...=0) and, respectively, after …=1) to have the sample statistics calculated.)

The answer you get (after pressing OK) is

```
 Null hypothesis: Difference of means = 0

Sample 1:
 n = 330, mean = 129,939, s.d. = 22,8908
 standard error of mean = 1,2601
 95% confidence interval for mean: 127,461 to 132,418
Sample 2:
 n = 379, mean = 127,942, s.d. = 19,254
 standard error of mean = 0,989011
 95% confidence interval for mean: 125,997 to 129,887

Test statistic: t(707) = (129,939 - 127,942)/1,58299 = 1,26182
Two-tailed p-value = 0,2074
(one-tailed = 0,1037)
```

[1] "More accurate" means that for nonsymmetric data (the definition of symmetry is given in Sec. 1.4) the sample median is generally closer to the true, or population, median than sample mean to the true mean. For symmetric data the accuracy is more or less the same in both cases.

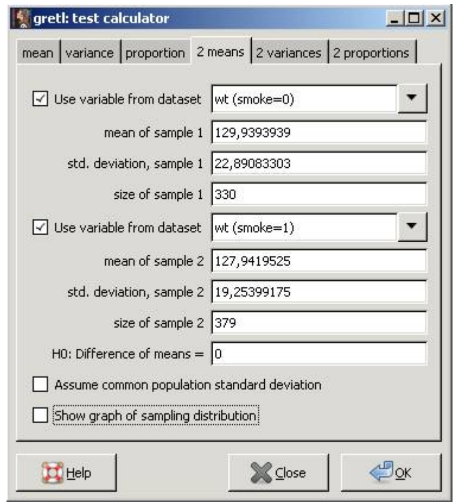



Fig. 1.3. To estimate the means for both subsamples

which means that the mean value in the non-smoking group `smoke=0`, namely 129,939 does not differ much from that in group `smoke=1`, namely 127,942. The question - is the difference significant or can it be explained just by sampling variations? – will be examined in Ch.2.

**1.1 exercise**. When your GRETL session is over, you can save your data in GRETL format (*.gdt): go to File| Save data, name the file parents.gdt, and close GRETL. Begin a new GRETL session and import fivenum.txt from ShortGRETLdata. Calculate the means and medians of both variables manually and with GRETL. Explain differences.

It is difficult to expect that one number, whether mean or median, will give a comprehensive description of a sample (and, ultimately, a population). Another number which helps to describe the population is the spread (or variation or dispersion) of the sample values around its mean. One of such characteristics is the sample **standard deviation** $s$ defined as $s = \sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 / (n-1)}$ [2]. In „good“[3] cases approximately 95% of all sample values belong to the interval $(\overline{x} - 2s, \overline{x} + 2s)$ (this is called a two sigma rule).

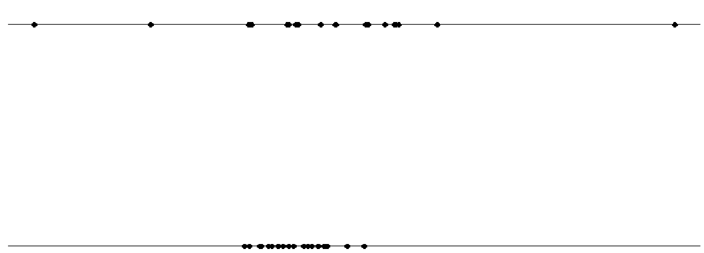


Fig. 1.4. The upper sample has a big standard deviation and the lower small.

[2] As $n \to \infty$, the sample standard deviation tends to that of population.
[3] “Good” means a variable whose distribution is close to normal or Gaussian (see the last section of this chapter).

So far we have investigated individual properties (i.e., average and spread values) of a single variable. A very important issue in statistics is to measure the strength of relationship between two or more variables. The respective most popular numeric characteristic for two variables is the **coefficient of correlation** $r$: it is always between -1 and +1; if it close to 0, the variables are „almost unrelated"; the linear relationship between the two variables is strong if $r$ is close to -1 or +1.

To evaluate $r$ between weight and height, open parents.gdt (go to File| Open data| User file...| parents.gdt), select `wt` and `ht`, right-click on your selection and choose Correlation matrix. You will get the following table:

```
corr(wt, ht) = 0,42216963
Under the null hypothesis of no correlation:
t(707) = 12,3829, with two-tailed p-value 0,0000
```

The correlation is rather strong (i.e., the (`ht`, `wh`) points on the scatter diagram in Fig. 1.9 do not digress far from a line). The positive value of $r$ indicates that higher values of `ht` induce higher values of `wt` (clearly, this is what we have expected.)

Numeric characteristics are very useful to get a general impression of our data set. However, no less useful are

## 1.4. Graphical Characteristics

We have already mentioned „symmetrically distributed data" which means that a variable takes values equally likely above or below the „average value". To get a feeling on the distribution of `ht`, check and right-click it, then choose Frequency distribution.

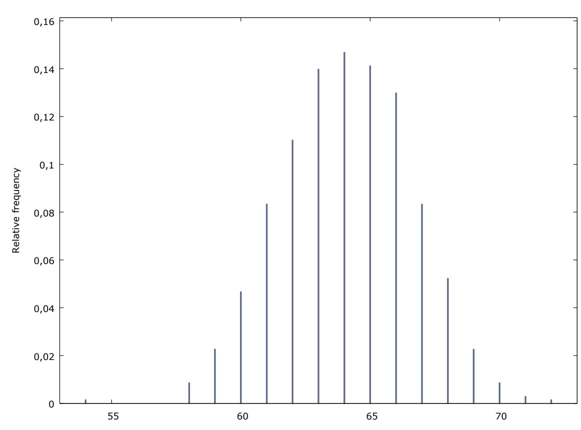



Fig. 1.5. **Histogram** of `ht` is almost symmetric and bell-shaped; however, as you will see later, not all bells are the same

In this figure, relative frequences are drawn. For example, the most frequent value of 64 repeats approximately in 0.15 or 15% of our observations (to get the exact number, repeat previous steps but uncheck the box „show plot" – you will get 14,67%). Note that in the case where a variable takes „many" values, the histogram is different, it is drawn for grouped values (see 1.2 exercise).

**1.2 exercise**. Draw histograms and print frequences of `wt`, `dwt`, and `dht`. How do you interpret the histogram of `smoke`?

Another graphical characteristic is the **boxplot**. The plot displays the distribution of a variable. The central box encloses the middle 50 percent of the data, i.e., it is bounded by the first and third sample quartiles[4]. The "whiskers" extend to the minimum and maximum values. A line is drawn across the box at the median and the "+" sign identifies the mean.

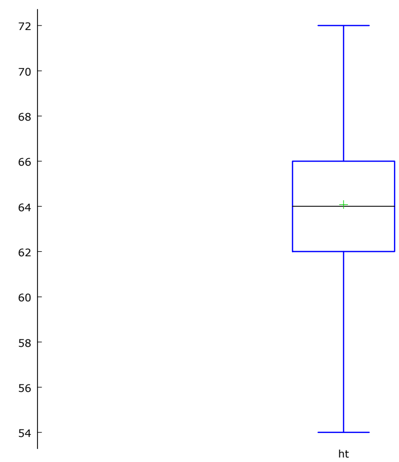



Fig. 1.6. Select, right-click `ht`, and press Boxplot; the boxplot of `ht` is almost symmetric with respect to its median, the mean and median of `ht` almost coincide

We already have got some idea about the differences in `wt` between smoking and non-smoking women. Another possibility is to draw two separate boxplots for each group. This is the first time when we shall use GRETL's **console** (command) window: click on console's icon (see bottom-left corner of GRETL). Type in `boxplot wt(smoke=0) wt(smoke=1)` and press Enter:

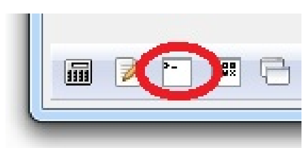


Fig. 1.7. Third from the left is the console's icon.

you will see Fig. 1.7 – two boxplots (except for maximum values) are very similar.

**1.3 exercise**. Open a new GRETL's window. Import (as time-series quarterly data, 1962:1-1995:4) caemp.txt file from ShortGRETLdata (this is seasonally adjusted Canadian index of employment) . Right-click the variable `caemp`, plot its graph, calculate its numeric and graphical characteristics. Present your findings in a MS Word file. ◀

[4] The first, second, and third sample quartiles of data values are the three points that divide the data set, rewritten in an ascending order, into four equal groups, each group comprising a quarter of the data. The second quartile is also called *median*. As always, when the sample size increases, the sample quartile tends to theoretical or population quartile. For example, if r.v. $U$ has a [1,5]-uniform distribution, then its third sample quartile will be close to the third theoretical quartile which equals 4 (because $P(U \le 4) = 3/4$ ).

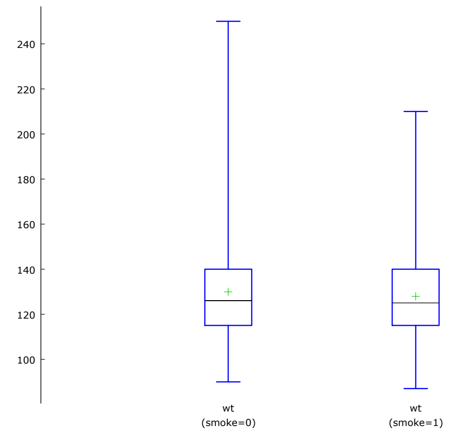



Fig. 1.8. `wt` boxplots for non-smoking and smoking women

So far we have investigated individual graphical properties of variables. A very convenient tool to investigate links between two variables is a scatter diagram. Open parents.gdt, select `wt` and `ht`, right-click and choose XY scatterplot.

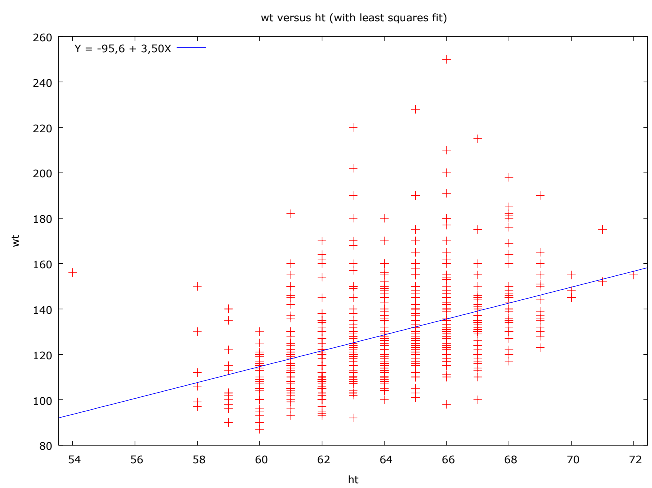



Fig. 1.9. Taller women weight more (a general trend is demonstrated by the blue (or *regression*) line)

**1.4 exercise**. Draw a scatter diagram of `dht` (x axis) and `dwt`. Draw a scatter diagram of `smoke` (x axis) and `wt`. Explain both plots. ◀

## 1.5. Normal Distribution

We have already met some variables whose distribution is bell shaped – see Fig. 1.5. In statistics, very important is the case where the bell is of some special form.

**1.1 definition.** We say that a variable $X$ has a normal or Gaussian distribution in the population if the chances of its values are described by the *density function* $f_X$ given by the formula $f_X(x) = \frac{1}{\sigma_X\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right), \quad -\infty < x < \infty$. Here $\mu_X$ is the mean of $X$ in the population and $\sigma_X$ is its standard deviation (if $\mu_X = 0$ and $\sigma_X = 1$, the distribution is called *standard* normal). The normal r.v. can take any value between $-\infty$ and $\infty$, but 95% of all the objects in the population belong to the interval $(\mu - 2\sigma, \mu + 2\sigma)$[5], thus, as a matter of fact, the normal r.v. is bounded[6].

The normal distribution is often used to describe, at least approximately, any variable that tends to cluster around the mean. For example, the heights of adult males in the United States are roughly normally distributed, with a mean of about 70 inches (1.8 m). Most men have a height close to the mean, though a small number of outliers have a height significantly above or below the mean. A histogram of male heights will appear similar to a bell curve, with the correspondence becoming closer if more data are used.

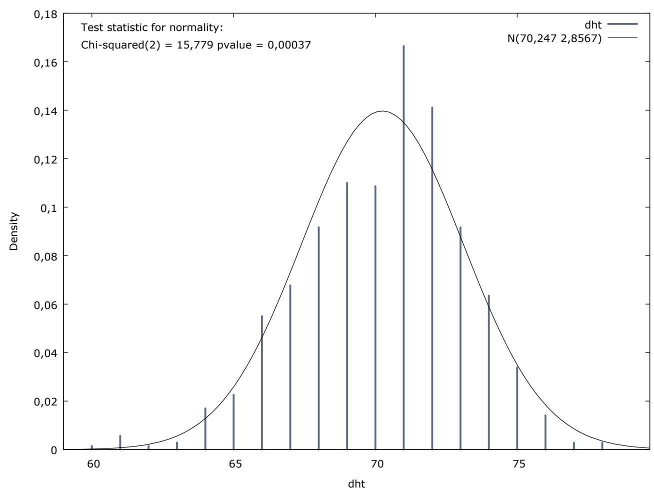



Fig. 1.10. The histogram of `dht` and respective normal density curve with mean $(\overline{X} =)$ 70.247 (inches) and standard deviation $(s =)$ 2.8567 (inches)

The histogram[7] in Fig. 1.10 does not follow the normal curve very well but we shall postpone the analysis of normality of `dht` till Ch. 2.

The normal distribution is very important because of the *Central Limit Theorem* which claims that whatever is the distribution of the summands, their sum (provided the number of summands is „big") will have a distribution close to normal. Thus, the sample mean $\overline{x} = (x_1 + ... + x_n)/n$ is a number in a concrete sample, but if one takes „many" similar samples, $\overline{x}$ will be different for different samples, thus it is a random variable, and its histogram will be close to normal.

[5] This is called the $2\sigma$ rule.

[6] The hight of adult men is satisfactory described by normal distrubution thus, in theory, it can take negative values. However, by the $2\sigma$ rule, to meet the man whose hight is outside the interval $(\mu - 2\sigma, \mu + 2\sigma)$ is little probable.

[7] In GRETL's window right-click on `dht`, choose Frequency distribution| Test against normal distribution.

**1.5 exercise**. Import the cross-sectional data contained in the GasCons.txt file from .../data folder. Find descriptive characteristics of the variable `cons`.

**1.6 exercise**. Go to File| Open data| Sample file...| Ramanathan, right-click data2-1, choose Info and then Open. Analyse and explain the data.

**1.7 exercise**. Go to File*Open data*Sample file...*Ramanathan, right-click data2-3, choose Info and then Open. The data consists of four *time series* (this means that any variable, say `unemp`, is measured at regular time intervals; in our case, they are measured once a year, therefore these are *annual* time series):

| | |
|---|---|
| `unemp` | civilian unemployment rate (%) |
| `cpi` | consumer price index (1982-84 = 100) |
| `infl` | percent change in cpi (inflation rate) |
| `wggr` | percent change in average weekly earnings (current dollars) |

1. Search (maybe with Google) for all the definitions of our four time series found with Info.

As you have just learned in 1, inflation is defined as

$$\mathtt{inf}_t = \frac{\mathtt{cpi}_t - \mathtt{cpi}_{t-1}}{\mathtt{cpi}_{t-1}} * 100\% \quad \text{(GRETL's syntax: } \frac{\mathtt{cpi} - \mathtt{cpi}(-1)}{\mathtt{cpi}(-1)} * 100\%)$$

which is numerically very close to a simpler expression of $\mathtt{inf2}_t = \log \mathtt{cpi}_t - \log \mathtt{cpi}_{t-1} = \Delta \log \mathtt{cpi}_t$. The inflation variable (noted in our data set as `infl`) is present in our data set but in order to practise in transforming our variables, recalculate it:

2. To create `inf2`, select `cpi`, go to Add and choose Log differences of selected variables – a new variable `ld_cpi` will appear in the list.
3. To create `inf`, we shall use gretl's console: type there `series inf=(cpi-cpi(-1))/cpi(-1)*100`
4. To compare the three inflations, select `infl`, `ld_cpi`, and `inf`, right-click on them and choose Time series plot.

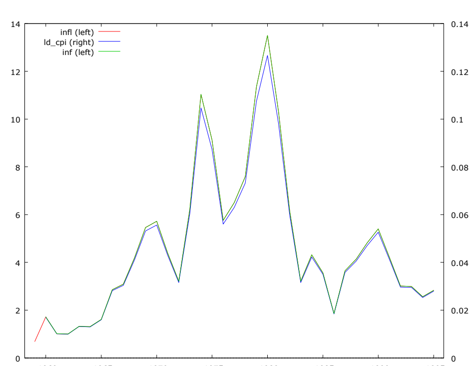



Fig. 1.11. `infl` and `inf` coincide whereas `ld_cpi` marginally differs.

Find the average inflation over these 37 years. Go to Sample*Set range... and choose the period from 1990 till 1995. What is the (average) inflation during these years? Do the same with unemployment.

5. To restore weekly earnings from `wggr`, use the following GRETL script:

```
series ear = 1
ear = (1+wggr(-1)/100)*ear(-1)
```

(can you restore and explain respective formula?). Plot the time series `ear`.

**1.8 exercise**. The data set pi2000.txt contains the first 2,000 digits of $\pi$. Draw its histogram. Is the distribution of the values of $\pi$, namely, 0, 1, …, 9 close to uniform, that is, are the relative frequences of each value equal to more or less the same (which?) number? What is the percentage of digits that are 3 or less? - use the line `scalar pp = sum(pi2000<=3)/2000`. Where is the answer placed?

**1.9 exercise**. The time variable in the nym.2002.txt data set contains the time to finish the 2002 New York City marathon for a random sample of the 1000 finishers (cross-sectional data).

1. What percent ran the race in under 3 hours (`time`=180)?
2. What is the time cutoff for the top 10%? The top 25%? (use Data| Sort data… by `time` or run the line `scalar q10 = quantile(time,0.10)` ).
3. What time cuts off the bottom 10%?
4. Do you expect the variables `age` and `time` to be symmetrically distributed? Make their histograms and describe the shape. Can you explain why the shape is as it is?

Later we shall develop methods to get a quantitative answer to the questions posed in the following exercises. At the moment, we shall analyze them „desciptively“.

**1.10 exercise**. A report was prepared on how to prevent aggressive driving and road rage. As described in the study, *road rage* is criminal behavior by motorists characterized by uncontrolled anger that results in violence or threatened violence on the road. One of the goals of the study was to determine when road rage occurs most often. The days on which 69 road rage incidents occurred are presented in the file rage.txt (M=Monday, Tu=Tuesday etc). Analyze the frequency distribution of rage. Is it true that number of incidents is (almost) the same in all days? Which day is the most unfortunate?

**1.11 exercise**. A study examined, among other issues, alcohol consumption patterns of U.S. adults by marital status. Data for marital status (`STATUS`) and number of drinks per month (`DRINKS`), based on the researchers' survey results, are provided in the marital.txt file. When imported to GRETL, the data will be recoded to the integer values 1, 2 etc and treated as discrete variables. a) Create a cross-tabulation table[8] of `STATUS` vs `DRINKS` (how do you think, do the variables interact?; i.e., does the distribution of the number of drinks depend on `STATUS`?) b) Go to Sample| Restrict, based on criterion… and type `STATUS`=1 etc; then analyze the Frequency distribution of `DRINKS` in all the stratas of `STATUS`; does the distribution of the number of drinks depend on `STATUS`?

[8] Select `STATUS` and `DRINKS` and go to View| Cross Tabulation.

**1.12 exercise**. Anthropologists are still trying to unravel the mystery of the origins of the Etruscan empire, a highly advanced Italic civilization formed around the eighth century B.C. in central Italy. Were they native to the Italian peninsula or, as many aspects of their civilization suggest, did they migrate from the East by land or sea? The maximum head breadth, in millimeters, of 70 modern Italian male skulls and that of 84 preserved Etruscan male skulls were analyzed to help researchers decide whether the Etruscans were native to Italy. The resulting data can be found in the etruscans.txt file. Analyze both variables, ITALIANS and ETRUSCANS (find their summary statistics and frequency distribution). Can we say that variables differ "considerably"? Why?

# 2. Hypothesis testing

Any assertion about population parameters is called a statistical hypothesis. We want to use our sample and decide whether this assertion is true. For example, is it true that the mean value $\mu$ of the population equals a concrete number $\mu_0$? (in short: is the main (or null) hypothesis $H_0 : \mu = \mu_0$ true?). The procecure for testing this assertion is called a statistical test, it is based on the discrepancy statistics. Most probably, you know that, according to the Law of Large Numbers, as the sample size increases, the sample mean $\bar{x}$ tends to the true mean of population $\mu$. Therefore, if our guess $\mu = \mu_0$ (or the null hypothesis) is true, the discrepancy between $\bar{x}$ and $\mu_0$ (in this case, it is simply the difference $|\bar{x} - \mu_0|$) should not be „very big" (in other words, if the difference is „big", the assumption $H_0$ is, most probably, false). There are many methods to define the term „big" but in statistics the ultimate measure of this discrepancy is the *p*-value of the test – if it is less than 0.05[1], the null is rejected and we accept the alternative assertion (or hypothesis) $H_1$ (it can be formulated as $\mu \neq \mu_0, \mu > \mu_0$ or $\mu < \mu_0$; the calculation of the *p*-value depends on the alternative but, in any case, you should not worry about it – the computer program will do it for yourself). In the case where we reject null (that is, the *p*-value is less than 0.05), we say that the true population mean differs *significantly* from our hypothetical value $\mu_0$, otherwise (that is, if the *p*-value is greater than 0.05 and, consequently, we accept null), we say that our data does not contradict our main assertion $\mu = \mu_0$.

The testing procedure of all the hypotheses is the same: we formulate the null $H_0$ and the alternative $H_1$; if the *p*-value of this test is greater than 0.05, we say that there is no reason to reject $H_0$ (in other words, we accept $H_0$); otherwise, we reject $H_0$ and accept $H_1$.

In the sequel, we shall examine some most popular tests.

## 2.1. Testing the mean (Student's $t$ - test)

Let us assume that we observe an (approximately) normal variable with unknown population mean $\mu$ and unknown[2] standard deviation $\sigma$. To test the hypothesis $H_0 : \mu = \mu_0$ (here $\mu_0$ is an arbitrary number of interest) with alternatives $H_1 : \mu \neq \mu_0$ or $H_1 : \mu > \mu_0$ or $H_1 : \mu < \mu_0$ we use the $t$ - test.

[1] To give the precise definition of the $p$-value is not so easy, therefore we skip it. In any case, you can treat the $p$-value as probability that the null is true – if the *p*-value is less than 0.05, it is little probable that $H_0$ is true, therefore we reject it (a popular saying says "if $p$ – value is low, $H_0$ must go"). 0.05 is the standart value of the *significance* of a test; the value can also be 0.10 or 0.01 or any other „small" number.

[2] Usually we have just a collection of numbers $x_1, ..., x_n$, thus it is natural to assume that we do not know neither the true mean $\mu$ nor standard deviation $\sigma$.

**2.1 example.** The manufacturer claims that the fuel city consumption of its car is 10.7 l/100 km. The owner registered the fuel consumption for two months (each day at 7pm), the results are in the file GasCons.txt. Test the manufacturer's claim. ◀◀

After importing the file, select the variable `cons`, go to Tools * Test statistic calculator * mean * check the box „Use variable from data set" * choose „H0:mean=" 10.7 * click OK. You get the following table:

```
Null hypothesis: population mean = 10,7
Sample size: n = 50
Sample mean = 11,4278, std. deviation = 2,74113
Test statistic: t(49) = (11,4278 - 10,7)/0,387654 = 1,87745
Two-tailed p-value = 0,06642
(one-tailed = 0,03321)
```

Can we accept the null $H_0: \mu = 10.7$ ? You have to take a look to the $p$-value[3]:

1. Assume that the alternative is $H_1: \mu \neq 10.7$; since the $p$-value is 0.066 (>0.05), we have no ground to reject the manufacturer's claim, i.e., we accept the null hypothesis.
2. Since the sample mean 11.4278 is greater than 10.7, the owner wants to test a more appropriate alternative, namely the one-sided hypothesis $H_1: \mu > 10.7$. The one-sided (or one-tailed) $p$-value equals (0.066/2=) 0.033. Since it is less than 0.05, we reject the manufacturer's claim (thus we have proved (with significance level 0.05) that the car consumes more than 10.7 l/100km). ◀◀

Sometimes, we do not know the whole sample, we are given only the sample mean, sample standard deviation, and the size of the sample (in our case, it is 11.4278, 2.74113, and 50). To estimate the $p$ - value, use the formula (see Statistics formula sheet, Test for population mean)

$$p - \text{value} = P(T_{n-1} > |\bar{x} - \mu_0| / (s / \sqrt{n}))$$

(here $\bar{x} - \mu_0$ stands for discrepancy and $T_k$ for the Student random variable with $k$ degrees of freedom). To calculate the $p$-value, go to GRETL's script window and run the following lines:

```
scalar n = 50
scalar s_mean = 11.4278
scalar hyp_mean = 10.7
scalar df = n-1
scalar s_sd = 2.7411
scalar t_stat = (s_mean - hyp_mean)/(s_sd/sqrt(n))
scalar pval = pvalue(t,df,abs(t_stat))   # t stands for Student's distribution
```

(in the Session icon view| Scalars window, you will find that `pval` equals 0.0332... )

**2.1 exercise.** Import the file parents.txt. Is it true that the mean value of the father's height `dht` is equal to 64,080[4] (inches)? Test the claim (i.e., find the $p$ – values) in two ways.

---

[3] In fact, $p$ - value gives the answer to the question: is the <u>sample</u> mean 11.4278 close enough to the hypothesized <u>population</u> mean 10.7?; since the $p$ - value is 0.033, the answer is „no".

[4] This is the mean value of mothers' height.

**2.2 exercise.** At Canon Food Corporation, it used to take an average of 90 minutes for new workers to learn a food processing job. Recently the company installed a new food processing machine. The supervisor at the company wants to find if the mean time taken by new workers to learn the food processing procedure on this new machine is different from 90 minutes. A sample of 20 workers showed that it took, on average, 85 minutes for them to learn the food processing procedure on the new machine. It is known that the learning times for all new workers are normally distributed with a sample standard deviation of 7 minutes (and the sample mean, of course, 85 minutes). Find the *p*-value for the test that the mean learning time for the food processing procedure on the new machine is i) different ii) less from 90 minutes.

## 2.2. Testing normality (chi-squared test)

In the previous section, we omitted an important step: recall that the $t$ – test is applicable to only (approximately) normal variables. Thus, is `cons` from GasCons.txt normal? To test this[5], import GasCons.txt, select `cons`, right-click on it and choose Frequency distribution * check “Test against normal distribution” * OK. You will get the histogram of `cons` (see Fig. 2.1) together with the output of the so-called chi-squared test on normality (see the top-left corner of the figure). The null in this case is $H_0 : \texttt{cons}\ \textit{is a normal variable}$ with the alternative $H_1 : \texttt{cons}\ \textit{is not a normal variable}$. If the null is true, the value of chi-squared(2) must be close to zero. Is the discrepancy[6] of 3.319 “big” enough to reject the null? The answer is given by the *p*-value: it equals 0.19019, i.e., it is greater than 0.05, thus we do not reject the normality assumption. In other words, the above performed *t*-test is a valid procedure.

**2.3 exercise.** Test the normality of `dht` and `dwt`.

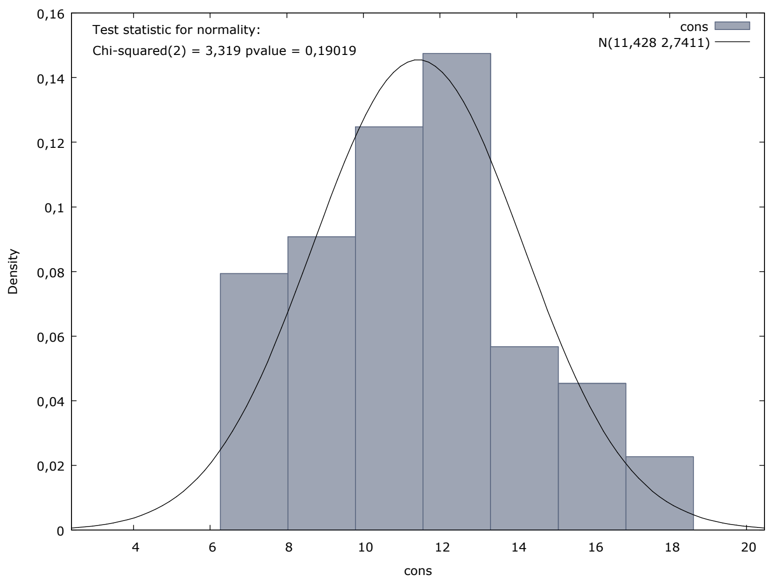



Fig. 2.1. The histogram of `cons`.

[5] There are many tests to test normality. Here we use the chi-squared test, other possibilities will be presented in 2.3 example below.
[6] Between the histogram and the normal density function.

**2.4 exercise.** Every year *Fortune* magazine publishes a list of the 100 best companies to work for (see http://money. cnn.com/magazines/fortune/best-companies/2012/full_list/ or the best2012.txt file).

```
                  company JobGrowth US_Employees
1                  Google        33        18500
2 BostonConsultingGroup        10         1958
3           SAS_Institute         8         6046
4      WegmansFoodMarkets         5        41717
5             EdwardJones         1        36937
6                  NetApp        30         6887
..............................................
```

- **a.** Construct histograms of the `JobGrowth`, `US_Employees` and their logarithms. Do the distributions of values appear to be reasonably symmetric? Test the data for normality.
- **b.** The `US_Employees` values are skewed to the high end. The logarithm transformation makes the distribution more nearly symmetric. A symmetric distribution is more appropriate to summarize with a mean and standard deviation.
- **c.** Another possibility to symmetrise the distribution is to remove, say, 40% of the highest values from the list. Do this and test again the `US_Employees` for normality.

## 2.3. Testing the equality of two means (Student's t-test)

Let us begin with an example.

**2.2 example.** In order to prove that a new medicine is effective in curing a disease, two groups of patients were formed: patients in the first group were taking the medicine and patients in the second one a harmless and ineffective substance (placebo). The data is given in the file medicine.txt:

```
time group
15   1
10   1
13   1
 7   1
 9   1
 8   1
21   1
 9   1
14   1
 8   1
10   1
19   1
10   1
11   1
 6   1
15   2
14   2
12   2
 8   2
14   2
 7   2
16   2
10   2
15   2
12   2
```

Here `time` is time until the patient recovers while `group` indicates whether the medicine or placebo was given to the patient.

Can we claim that the medicine is more effective than placebo? To get some ideas about the problem, draw two boxplots (one for each group)[7]. In Fig.2.2, it is easy to see that the median of the time necessary to recover for the patients taking placebo is considerably greater[8] than that of patients taking the medicine (that is the medicine is to some extent effective). But can we substantiate our claim of medicine's superiority?

[7] In GRETL's console type `boxplot time(group=1) time(group=2).`
[8] Note that difference between means is not that great.

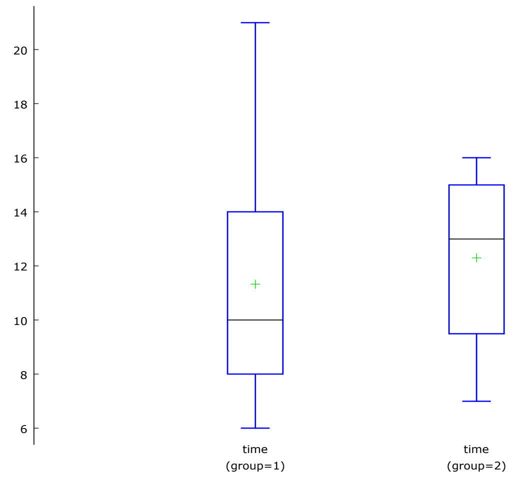



Fig. 2.2. The boxplot of `time` for `group`=1 (left) and `group`=2 (right)

To formalize our problem, let $X_1,...,X_n$ be normal variables with unknown mean $\mu_X$ and unknown standard deviation $\sigma$ and independent $Y_1,...,Y_m$[9] be normal variables with unknown mean $\mu_Y$ and unknown standard deviation $\sigma$ (note that i) we assume equal standard deviations in both populations and ii) $n$ not necessarily equals $m$). We use $t$- test to test the null (of equality of means) $H_0: \mu_X = \mu_Y$ versus the alternative $H_1: \mu_X \neq \mu_Y$ or $H_1: \mu_X > \mu_Y$ or $H_1: \mu_X < \mu_Y$.

**Remark.** Before applying this test one has to test equality of variances of $X$ and $Y$ (if variances are not equal, GRETL will present only an approximate $p$- value of the test). ◀◀

Thus, in order to prove that medicine is superior to placebo we shall apply $t$- test. But first we test the equality of variances: go to Tools * Test statistic calculator * 2 variances * check the first box, print `time(group=1)` and press Enter, then check the second box, print `time(group=2)` and press Enter * OK. You get the following table:

```
Null hypothesis: The population variances are equal
Sample 1:
n = 15, variance = 18,6667
Sample 2:
n = 10, variance = 9,56667
Test statistic: F(14, 9) = 1,95122
Two-tailed p-value = 0,3153
(one-tailed = 0,1577)
```

The most important number here is the $p$ - value `0.3153`. Since it is greater than 0.05, we have no ground to reject the null $H_0: \sigma_X^2 = \sigma_Y^2$[10]. Thus, we can apply the $t$ - test: go to Tools * Test statistic

[9] In our case, $X$ is `time` in the first group and $Y$ is `time` in the second group.

[10] Note that $s_X^2 =$ `18.6667` and $s_Y^2 =$ `9.56667`, i.e., the first sample variance is twice as big; nevertheless, we proved that this does not contradict (with any reasonable significance level) the null $\sigma_X^2 = \sigma_Y^2$.

calculator * 2 means * check the first box and print `time(group=1)` and press Enter * check the second box and print `time(group=2)` and press Enter * OK. You will obtain the following table:

```
Null hypothesis: Difference of means = 0

Sample 1:
 n = 15, mean = 11,3333, s.d. = 4,32049
 standard error of mean = 1,11555
 95% confidence interval for mean: 8,94072 to 13,7259

Sample 2:
 n = 10, mean = 12,3, s.d. = 3,093
 standard error of mean = 0,978093
 95% confidence interval for mean: 10,0874 to 14,5126

Test statistic: t(23) = (11,3333 - 12,3)/1,58671 = -0,609229
Two-tailed p-value = 0,5483
(one-tailed = 0,2742)
```

To prove the superiority of the medicine, choose the one-sided alternative $H_1 : \mu_X < \mu_Y$ [11] and a respective one-tailed $p$ - value 0.2742. It is greater than 0.05, therefore we do not reject $H_0 : \mu_1 = \mu_2$. The bottom line: the experiment failed to prove the superiority of new medicine. ◀◀

What to do if variances in the two groups are not equal? To get an approximate $p$-value, uncheck the "Assume common population standard deviation" box (in our case, you will get practically the same result).

Now suppose that we do not have the sample given in medicine.txt, but we only know that $n1 = 15$, $mean1 = 11.333, sd1 = 4.320$ and $n2 = 10, mean2 = 12.3, sd2 = 3.093$. To find the $p-$value, assuming that variances are equal, we use the formula[12] (cf. the Two sample *t*-test in the last chapter of these Notes)

$$p - \text{value} = P(T_{n_1+n_2-2} > | \bar{x}_1 - \bar{x}_2 | /(s / \sqrt{1/n_1 + 1/n_2}))$$

where $s = \sqrt{\dfrac{(n_1 - 1)s_{X_1}^2 + (n_2 - 1)s_{X_2}^2}{n_1 + n_2 - 2}}$ is an estimate of the pooled standard deviation and $| \bar{x}_1 - \bar{x}_2 |$ stands for discrepancy; in GRETL, run the following lines:

```
scalar n1 = 15
scalar mean1 = 11.333
scalar sd1 = 4.320
scalar n2 = 10
scalar mean2 = 12.3
scalar sd2 = 3.093
#############
scalar df = n1+n2-2
```

[11] This means that the average time to recover when taking the medicine is shorter than the time with placebo.

[12] This is the last formula for the calculation of the $p$ - value in these notes. To find relevant formulas for other tests, use any more advanced textbook or consult the last chapter of the present notes (Statistics formula sheet).

```
scalar st_dev = sqrt(((n1-1)*sd1^2+(n2-1)*sd2^2)/df)*sqrt(1/n1+1/n2)
scalar t_stat = (mean1-mean2)/st_dev
scalar pval = pvalue(t,df,abs(t_stat))
```

In the script output and also in the Scalars windows you see `pval=`0.2741.

**2.3 example.** Prior to our analysis, we had to test the normality of `time`. We already know how to do this but in our case we have an additional complication: we have to test normality in each group separately. To choose the first group, in GRETL menu line choose Sample * Restrict, based on criterion..., then type `group==1`[13]. Now, to test normality in the first group, select `time` and choose Variable * Normality test – you will get the following table:

```
Test for normality of time:
Doornik-Hansen test = 4,5439, with p-value 0,103111
Shapiro-Wilk W = 0,893866, with p-value 0,0767465
Lilliefors test = 0,22119, with p-value ~= 0,05
Jarque-Bera test = 2,4849, with p-value 0,288676
```

As you can see, there are many tests to test normality. Almost all of them (except, maybe, Lilliefors test) do not reject normality (because all the $p$ - values are greater than 0.05).

To test normality in the second group, repeat the above procedure, but now type `group==2` and check „replace current restriction“ box.

```
Test for normality of time:
Doornik-Hansen test = 1,90405, with p-value 0,385958
Shapiro-Wilk W = 0,912493, with p-value 0,298558
Lilliefors test = 0,208713, with p-value ~= 0,24
Jarque-Bera test = 0,949123, with p-value 0,622158
```

Again, we do not reject normality in the second group. Thus, we can trust our $t$ - testing result.

**2.5 exercise.** The file C:\ShortIntro\hsb.txt contains the high school survey data. Import the data to GRETL (with separator for data columns: space). Among other variables, it contains `SEX`, `RACE`, `SES`, `MATH`, and `WRTG`.

`SEX`
1 MALE
2 FEMALE

`RACE`
1 HISPANIC
2 ASIAN
3 BLACK
4 WHITE

`SES` SOCIO-ECONOMIC STATUS
1 LOWER
2 MIDDLE
3 UPPER

`MATH` MATH T-SCORE
`WRTG` WRITING T-SCORE

[13] Note the change in the line right below the GRETL window.

Test the claim that asians are very clever in math. In order to do this, combine asians in one group with

```
series asian = ! (RACE==1||RACE==3||RACE==4)
```

Here || means OR, `(RACE==1||RACE==3||RACE==4)` equals 1 if the `RACE` is either HISPANIC, BLACK or WHITE, ! is a negation operator (thus, `asian` equals 1 if the student is ASIAN[14] and 0 otherwise.)

Draw two boxplots of `MATH` for asians and non-asians, test the equality of variances in these two groups, test the equality of means in these two groups. What is your conclusion? Also test the hypothesis that girls are better than boys in writing. ◄◄

**2.6 exercise.** Let $n_1 = 7, \overline{x}_1 = 185.07, s_{X_1}^2 = 443.80, n_2 = 8, \overline{x}_2 = 211.4, s_{X_2}^2 = 101.01$. Assuming equal variances in the (normal) populations, test the hypothesis $H_0 : \mu_1 = \mu_2$. ◄◄

One special case of the two-sample $t$- test is called a paired $t$- test. A typical situation situation where the test is applied is „before and after". In this case of <u>dependent</u> observations, not the individual $X$ and $Y$ observations are used, but, instead, the differences $D_i = X_i - Y_i$, $i = 1,...,n$, are formed and a one-sample null $H_0 : \mu_D = 0$ versus two- or one-sided alternative is tested[15]. For example, to compare a peak expiratory flow rate `PEFR` before and after a walk on a cold winter's day for a random sample of 9 asthmatics, an experimental data was collected (see pefr.txt file).

**2.7 exercise.** Explore the data set pefr.txt. Test the null $H_0 : \mu_D = 0$ versus the alternative $H_1 : \mu_D > 0$[16].

**2.8 exercise.** Import the exercise.txt file. Is it true that 1) the increase of the pulse `PULSE_2` – `PULSE_1` after the 1 mile run and 2) percentage increase of the pulse (`PULSE_2` – `PULSE_1`)/`PULSE_1` differs for men and women? Do these differences depend on `SMOKE`?

**2.9 exercise.** In 1.12 exercise, we introduced the etruscans.txt file. Pedantically test whether variables ITALIANS and ETRUSCANS have the same mean.

## 2.4. Testing hypothesis about proportion

Assume that we observe a variable taking only two values, 1 and 0 (say, success and failure). The ratio of the number of ones with the number of observations is called a relative frequency or the share of successes in the <u>sample</u>. The null hypothesis about the share of successes in the <u>population</u> $H_0 : p = p_0$ ($p_0, 0 < p_0 < 1$, is the number of interest) versus alternative $H_1 : p \neq p_0$ or $H_1 : p < p_0$ or $H_1 : p > p_0$ is tested with the proportion test.

---

[14] Frankly speaking, asians can be extracted with a much simpler command: `series asian = RACE = = 2`. The above example is to demonstrate the use of Boolen (logical) operators.

[15] Under assumption, that differences $D_i$ are from normal population.

[16] To create differences, use `series D = Before - After`.

**2.4 example.** The results of a pre-elective 1000 voters survey A2 are presented in survA2.xls (here 1 means that a responded intends to vote for the A party). Earlier survey established that 15% of the voters intended to vote for A. Has the popularity of A changed?

To answer the question, import the file survA2.xls and click on Tools * Test statistic calculator * Proportion, type in „yes2“ and H0: proportion = 0.15. The answer is as follows:

```
Null hypothesis: population proportion = 0,15
Sample size: n = 1000
Sample proportion = 0,136
Test statistic: z = (0,136 - 0,15)/0,0112916 = -1,23986
Two-tailed p-value = 0,215
(one-tailed = 0,1075)
```

As all the $p$ - values are greater than 0.05, whatever is the alternative, we do not reject $H_0$, i.e., despite the frequency decrease in the given sample, the popularity of the party in the whole population of voters remains the same (with the significance level 95%). ◀◀

So far we have discussed the one proportion case. More often we confront with two or more[17] proportions. To formalize the problem, assume that $X$ was observed N1 times with M1 successes[18] and variable $Y$ N2 times with M2 successes. We want to test the null $H_0: p_X = p_Y$ of the equality of proportions in the two populations with relevant alternative[19].

**2.5 example.** In fact, we know not only the previous proportion 0.15, but also the whole file of the previous survey A1. This additional information allows us to recalculate the $p$ - value. Note that the 2 proportions test in GRETL is arranged somewhat differently from other tests. We do not need the samples themselves but we just need both relative frequences and sizes of respective samples. To append GRETL‘s workfile with a new variable, choose File * Append data * Excel... * .../ShortIntro/survA1.xls. Now right-click on `yes1` and choose Descriptive statistics – you will see that we have 800 valid observations (other 200 are missing values). Go to Tools * Test statistic calculator * 2 proportions and fill the boxes, respectively, with 0.15, 800, 0.136, 1000.

```
Null hypothesis: the population proportions are equal

Sample 1:
n = 800, proportion = 0,15

Sample 2:
n = 1000, proportion = 0,136

Test statistic: z = (0,15 - 0,136) / 0,0165677 = 0,845017
Two-tailed p-value = 0,3981
(one-tailed = 0,1991)
```

[17] GRETL allows to examine only a two proportion case.
[18] Thus, M1/N1 is the share of successes in the sample.
[19] According to the Law of Large Numbers, the relative frequency M1/N1, tends, as $N1 \to \infty$, to the proportion of successes in the population, namely, $p_X$. The test examines whether the relative frequences M1/N1 and M2/N2 are close enough to accept the null hypothesis $p_X = p_Y$. The answer to this question is given by the $p$ - value.

Thus, we get the same answer as before – there is no ground to reject null, the voters' attitude did not change between these two surveys. ◄◄

**2.10 exercise.** A swimming school wants to determine whether a recently hired instructor is working out. Sixteen out of 25 of instructor A's students passed the lifeguard certification test on the first try. In comparison, 57 out of 72 of more experienced instructor B's students passed the test on the first try. Is instructor A's success rate worse than instructor B's?

**2.11 exercise.** Suppose the Luna Drug Company develops a new drug, designed to prevent colds. The company states that the drug is equally effective for men and women. To test this claim, they choose a a simple random sample of 100 women and 200 men. At the end of the study, 38% of the women caught a cold; and 51% of the men caught a cold. Based on these findings, can we reject the company's claim that the drug is equally effective for men and women?

## 2.5. Testing the equality of many means (ANOVA test)

In Section 2.3 we tested the equality of means of two independent populations. Unfortunately, the $t$ - test we used there cannot be applied in the case of more than two populations. To formalize the problem, assume that we have $G > 2$ independent groups of normal observations $X_{11}, X_{12},..., X_{1n_1} \sim N(\mu_{X_1}, \sigma^2)$, …, $X_{G1}, X_{G2},..., X_{1n_G} \sim N(\mu_{X_G}, \sigma^2)$ with, possibly, different means and the same variance $\sigma^2$. To test the null $H_0 : \mu_{X_1} = ... = \mu_{X_G}$ with the alternative $H_1$ : *at least one mean is different from the others* we use the ANOVA test (ANOVA means ANalysis Of VAriance but the term might be misleading – we are interested in means, not in variances).

**2.6 example.** Import the fecun.txt file where you will find two columns of data:

| | |
|---|---|
| `fecun` | fecundity of a fruit fly per day |
| `strain` | genetic line (three levels) |

To get an impression of our data, draw a boxplot of `fecun` for every group (in GRETL's console type boxplot fecun(strain=1) fecun(strain=2) fecun(strain=3), see Fig. 2.3). One can see that 1) sample means (marked with +) are rather different, thus population means in these three groups are, probably, also different, 2) the spread of values in each group (the width of boxes) is similar, thus variances in groups are, probably, equal, and 3) distributions are rather symmetric so they, probably, do not differ much from normal.

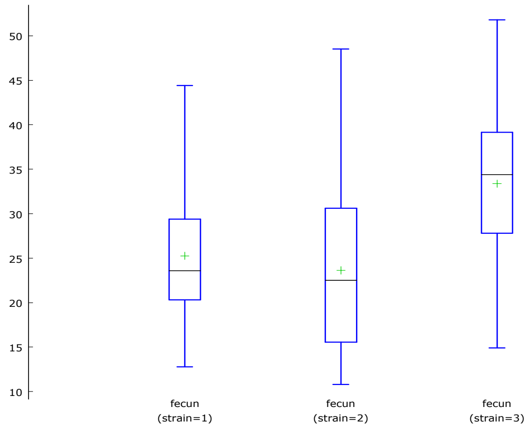



Fig. 2.3. Boxplots of `fecun` for each group

Thus, it seems likely that the ANOVA assumptions of equal variances in groups and normality are satisfied[20], therefore we apply the ANOVA test: in GRETL, go to Model * Other linear models * ANOVA… * Response variable `fecun` and Treatment variable `strain` * OK:

```
Analysis of Variance, response = fecun, treatment = strain:

                     Sum of squares       df      Mean square

  Treatment                 1362,21        2          681,106
  Residual                  5659,02       72          78,5975
  Total                     7021,23       74          94,8815
  F(2, 72) = 681,106 / 78,5975 = 8,66574 [p-value 0,0004]
  Level           n        mean      std. dev

  1              25      25,256        7,7724
  2              25      23,628        9,7685
  3              25      33,372        8,9420

  Grand mean = 27,4187
```

As always, the answer is given by the $p$ - value: since 0.0004<0.05, we reject the null of equal means.

**2.12 exercise.** A manufacturer of paper used for making grocery bags is interested in improving the tensile strength of the product. Product engineering thinks that tensile strength is a function of the hardwood concentration in the pulp and that the range of hardwood concentrations of practical interest is between 5% and 20%. A team of engineers responsible for the study decides to investigate four levels of hardwood concentration: 5%, 10%, 15%, and 20%. They decide to make up six test specimens at each concentration level. All 24 specimens are tested on a laboratory tensile tester. The data from this experiment are presented in tens_strength.xls. Do you accept the null that the strength does not depend on concentration in the pulp?

[20] Strictly speaking, we ought to test these hypotheses.

## 2.6. Significance test for correlation coefficient

After calculating a sample correlation coefficient $r$ (see p. 1-4) between $X$ and $Y$, it is usually reasonable to check its significance, i.e., to test the null $H_0 : \rho = 0$ with alternative $H_1 : \rho \neq 0$ (here $\rho$ is the cofficient of correlation in the population). Note that the below presented procedure requires both the $X$ and $Y$ samples to be normal. If samples differ slightly from normal distribution, this test is applicable, but its results will be not accurate. As deviation increases, the results become less credible.

To test the significance for the strength of (linear) relationship between weight `wt` and height `ht` in parents.txt, recall the procedure of p. 1-4:

```
corr(wt, ht) = 0,42216963
Under the null hypothesis of no correlation:
t(707) = 12,3829, with two-tailed p-value 0,0000
```

Since the $p$ - value is considerably less than 0.05, we reject the null, i.e., `wt` and `ht` are obviously correlated.

**2.13 exercise.** The file hsb.txt contains, among others, the variables MATH (i.e., math t-score) and WRTG (i.e., writing t-score). Plot the scatter diagram of these variables. Guess if the correlation coefficient equals zero. Check your claim.

## 2.7. Test for Independence

Categorical variables require a different approach, since they are less amenable to graphical analyses and because common statistical summaries, such as mean and standard deviation, are inapplicable (we use instead tabular descriptions). If we observe two group (or nominal) variables, we do not use the correlation test to check for „no relationship“. That test is replaced by the test for independence which examines whether the row and column variables are independent[21] of each other. This is the null hypothesis. Note that the procedure produces the correct $p$ - value only if the expected frequencies of each category is at least 5.

**2.7 example.** The data file grades.txt contains students grade and their grade in their previous class (graded on American A-F scale).

| | |
|---|---|
| `prev` | The grade in the previous class in the subject matter |
| `grade` | The grade in the current class |

The American style grades are A+, A, A-, B+, B, B-, C+, C, C-, D, and F. These symbols are not numbers, the variables `prev` and `grade` are called group variables. To analyze them, select both these variables and go to View * Cross Tabulation and check Show zeros explicitly. You will get the following table:

[21] We test the probabilistic independence, but it is close to everyday concept of independence.

**Table 2.1**

`Cross-tabulation of prev (rows) against grade (columns)`

| | [ 1] | [ 2] | [ 3] | [ 4] | [ 5] | [ 6] | [ 7] | [ 8] | [ 9] | TOT. |
|---|---|---|---|---|---|---|---|---|---|---|
| [ 1] | 2 | 2 | 1 | 1 | 0 | 1 | 0 | 2 | 0 | 9 |
| [ 2] | 1 | 1 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 5 |
| [ 3] | 0 | 0 | 11 | 1 | 1 | 4 | 3 | 1 | 1 | 22 |
| [ 4] | 1 | 3 | 0 | 4 | 15 | 2 | 3 | 0 | 0 | 28 |
| [ 5] | 0 | 0 | 7 | 1 | 1 | 9 | 5 | 1 | 3 | 27 |
| [ 6] | 1 | 1 | 2 | 4 | 0 | 0 | 3 | 3 | 1 | 15 |
| [ 7] | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 3 |
| [ 8] | 0 | 0 | 1 | 1 | 0 | 3 | 4 | 0 | 0 | 9 |
| [ 9] | 0 | 1 | 0 | 2 | 0 | 0 | 1 | 0 | 0 | 4 |
| TOTAL | 5 | 9 | 22 | 14 | 21 | 19 | 20 | 7 | 5 | 122 |

```
Pearson chi-square test = 137,265 (64 df, p-value = 2,90446e-007)
Warning: Less than of 80% of cells had expected values of 5 or greater.
```

Note that when importing GRETL changed grades in `prev`: B+ to 1, A- to 2, A to 4 etc and also in `grade`: B+ to 1, A- to 2, B to 4 etc. Note that these 1, 2 etc should be treated not as numbers but just as groups‘ names. On the contrary, inside the table we have numbers, namely, 2 in the table shows the number of occurencies or frequency of the pair (1,1) or, more specifically, (B+,B+). We guess that students‘ abilities do not change in one year, therefore if a student was graded B+ in previous class, his grade next year will be more or less the same. In any case, we expect that the variables `prev` and `grade` are dependent. As the $p$ - value of the independence test is less than 0.05, we reject the null of independence which proves our expectation. Note that in many cells expected values are less than 5, therefore our conclusion could be called doubtful. On the other hand, the $p$ - value is much less than 0.05, therefore it is little probable that more appropriate tests of independence will revise our conclusion.

**2.14 exercise.** The file spouse.txt gives the ages at marriage for a sample of 100 couples that applied for marriage licences:

| | |
|---|---|
| `COUPLE` | observation number |
| `HUSBAND` | husband‘s age |
| `WIFE` | wife‘s age |

Draw a scatter diagram. How do you think: are `HUSBAND` and `WIFE` related? Test the null $H_0$: the correlation coefficient between these two variables is zero. Our data are numeric variables but quite often we encounter grouped data. If you type in Add * Define new variable... `series hu = (HUSBAND>=30) + (HUSBAND>=50)` (and similarly `series wi = (WIFE>=30) + (WIFE>=50)`), you will recode the original values of `HUSBAND` with 0 for age <30, 1 for $30 \le$ HUSBAND <50 and 2 otherwise. Test independence of discrete variables `hu` and `wi`. ◀◀

**2.13 exercise**. Are the variables `age` and `time` in the nym2002.txt data set correlated? Plot their scatter diagram. Test respective hypothesis.

**2.14 exercise**. The data set normtemp.txt contains body measurements for 130 healthy, randomly selected individuals (this data set was used to investigate the claim that “normal” temperature is 98.6$^0$F degrees. The variable `temperature` contains normal body temperature, the varia-

ble `gender` equals 1 for males and 2 for females, and `hr` stands for the resting heart rate. i) Test the hypothesis that the mean temperature in the whole sample is 98.2, ii) Perform a two-sample test to see whether the male and female population means are equal, iii) Test whether cor(temperature,hr)=0, iv) Does the `hr` for males and females differ? Is the difference significant?

**2.15 exercise**. In 1.11 exercise, we have analysed relationship between marital STATUS and number of DRINKS per month. Are these nominal variables independent?

**2.16 exercise**. Philosophical and health issues are prompting an increasing number of Taiwanese to switch to a vegetarian lifestyle. A sudy compared the daily intake of nutrients by vegetarians and omnivores living in Taiwan. Among the nutrients considered was protein. Too little protein stunts growth and interferes with all bodily functions; too much protein puts a strain on the kidneys, can cause diarrhea and dehydration, and can leach calcium from bones and teeth. The data in vegetarians.txt give the daily protein intake, in grams, by samples of 51 female vegetarians and 53 female omnivores.

**a.** Obtain boxplots for the protein-intake data in the two samples.
**b.** Use the boxplots obtained in part (a) to compare the protein intakes of the females in the two samples, paying special attention to center and variation.
**c.** Obtain a histogram of the data and use it to assess the (approximate) normality of the variable under consideration.
**d.** What about the equality of variances in these two samples?
**e**. Do the data provide sufficient evidence to conclude that the mean daily protein intakes of female vegetarians and female omnivores differ? Perform the required hypothesis test at the 1% significance level.

**2.17 exercise**. The file une.txt contains state-wise unemployment rate in the U.S in 2003, 2004, 2005, and 2008 years (the data is from www.infoplease.com/ipa/A0931330.html).
**a**. Test whether `un` has a normal distribution each year.
**b**. Explain what the following script does and comment the printout:

```
summary un --simple --by=yrs
anova un yrs
boxplot un yrs --factorized --output=display
```

# 3. Regression Analysis

In this chapter, we deal with two sets of data where interest lies in either examining how one variable relates to a number of others or in predicting one variable from others. Multiple linear regression is a method of analysis for assessing the strength of the relationship between each of a set of explanatory variables, and a single response variable. When only a single explanatory variable is involved, we have what is generally referred to as *simple* linear regression.

Applying multiple regression analysis to a set of data results in what are known as regression coefficients, one for each explanatory variable. These coefficients give the estimated change in the response variable associated with a unit change in the corresponding explanatory variable, conditional on the other explanatory variables remaining constant (this is called a ceteris paribus condition). The fit of a multiple regression model can be judged in various ways, for example, calculation of the multiple correlation coefficient or by the examination of residuals, each of which will be illustrated later.

Later in this chapter, we shall study the data set CPS1985.txt:

```
ID wage education experience age ethnicity region gender occupation        sector union married
1  4.95          9         42  57      cauc  other female     worker manufacturing    no     yes
2  6.67         12          1  19      cauc  other   male     worker manufacturing    no      no
3  4.00         12          4  22      cauc  other   male     worker         other    no      no
4  7.50         12         17  35      cauc  other   male     worker         other    no     yes
5 13.07         13          9  28      cauc  other   male     worker         other   yes      no
...................................................................................
```

where

```
wage         Wage (in dollars per hour)
education    Number of years of education
experience   Number of years of potential work experience (age - education - 6)
age          Age in years
ethnicity    "cauc" (→1), "hispanic"(→3), or "other"(→2)
region       Does the individual live in the South? (South → 2, Other → 1)
gender       Gender (Female → 1, Male → 2)
occupation   Factor with levels "worker" (tradesperson or assembly line worker),
             „technical" (technical or professional worker), "services" (service
             worker), "office" (office and clerical worker), "sales" (sales work-
             er), "management" (management and administration)
sector       "manufacturing" (manufacturing or mining), "construction", "other"
union        Does the individual work on a union job?
married      Is the individual married?
```

Note that the string (or nominal) variables when imported to GRETL are recoded to numbers (for example, ethnicity takes on values `cauc`, `hispanic`, `other`, therefore they will be recoded to numbers 1, 3, 2):

```
One or more non-numeric variables were found.
Gretl cannot handle such variables directly, so they
have been given numeric codes as follows.

String code table for variable 6 (ethnicity):
  1 = 'cauc'
  2 = 'other'
  3 = 'hispanic'
```

etc. We want to understand how the variable wage relates to other variables and also to get some numerical characteristics of the goodness-of-fit of a model.

## 3.1. Simple Linear Regression

It is clear that wage depends on education, experience, age etc:

$$\mathtt{wage} = f(\mathtt{education}, \mathtt{experience}, \mathtt{age}, ...)$$

Many factors affecting wage are given in this data set. However, the most important, namely, abilities, organizational skills, and ambitions, are not present there (and, obviously, they cannot be easily quantified). After denoting these and other missing factors by $\varepsilon$ (it is called an *error* term), the above formula can be corrected to

$$\mathtt{wage} = f(\mathtt{education}, \mathtt{experience}, \mathtt{age}, ...) + \varepsilon\,;$$

here $f$ is called a regression function. The simplest form of the function is the linear one:

$$f(\mathtt{education}, \mathtt{experience}, \mathtt{age}, ...) = \alpha + \beta_1 \mathtt{education} + \beta_2 \mathtt{experience} + \beta_3 \mathtt{age} + ...$$

However, we do not know the true values of the coefficients; the most popular, namely, the (ordinary) least squares (OLS) method to estimate its unknown coefficients $\alpha, \beta_1, \beta_2, ...$ will be discussed later.

We begin this chapter with a *simple regression* case – we analyse the dependence of wage on only one variable, say, experience (that is, education, age and other variables will be included into the error term $\varepsilon$):

$$\mathtt{wage} = f(\mathtt{experience}) + \varepsilon\,.$$

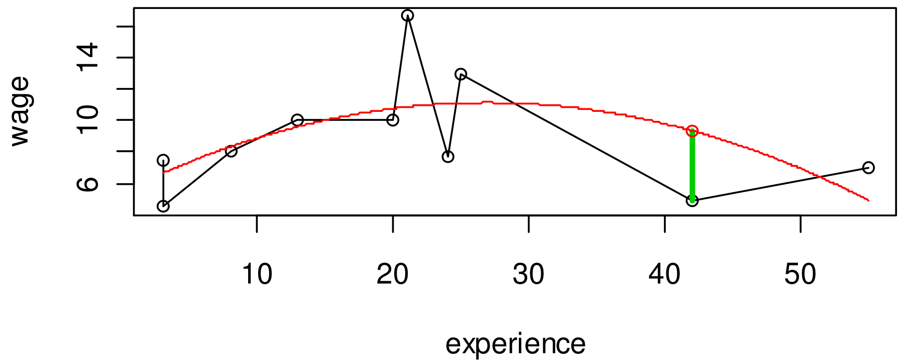



Fig. 3.1. Ten randomly selected cases from CPS1985.txt (black points) and a parabolic regression curve (red). The length of green segment denotes discrepancy between the parabolic regression curve and the observed value of wage (i.e., it represents one of many *error terms* $\varepsilon_i$ ).

In Fig. 3.1 one can see an `experience` - `wage` scatter diagram based on ten randomly selected cases of CPS1985.txt. A red regression curve, drawn through the „middle of the cloud" of our points, is designed to demonstrate a „general" dependence between `wage` and `experience`, it has to be „regular" (or smooth or „nice looking") and reflect the economic logic[1]. Generally speaking, if we have two models, we choose the one with smaller residuals[2] $\hat{\varepsilon}_i$ (more specifically, the one with the smaller sum $\sum_{i=1}^{n}\hat{\varepsilon}_i^2$). Of course, the residuals would have been the smallest (equal to zeros) if we had taken $f$ to be a black broken line in Fig. 3.1, however, it is definitely not a „nice" curve.

The most simple, though not always the most appropriate candidate for a regression curve is a straight line: $y = f_1(x) = \alpha + \beta_1 x$ (the coefficient $\alpha$ is called an *intercept* while $\beta_1$ the *slope* of the regression line); a bit more complicated model is given by the quadratic curve or parabola: $y = f_2(x) = \alpha + \beta_1 x + \beta_2 x^2$. In what follows, we use the so-called *ordinary least squares* (OLS) method to find the *estimates* of these coefficients. The OLS estimates of the coefficients of the model $y = \alpha + \beta_1 x + \varepsilon$ are the numbers $\hat{\alpha}$ and $\hat{\beta}_1$ such that the sum of squared residuals $SSR = SSR(\hat{a},\hat{b}_1) = \sum_{i=1}^{n}(y_i - (\hat{a} + \hat{b}_1 x_i))^2$ is the least possible[3]. To find the estimates in GRETL, after importing CPS1985.txt, we start with a linear model and go to Model * Ordinary Least Squares... * move `wage` to Dependent variable box and `experience` to Independent variables box * OK. We get the following Model 1:

```
Model 1: OLS, using observations 1-533
Dependent variable: wage

                 coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------------
  const          8,38474       0,389135     21,55      1,83e-074 ***
  experience     0,0362978     0,0179370     2,024     0,0435    **

R-squared              0,007653   Adjusted R-squared   0,005784
F(1, 531)              4,095074   P-value(F)           0,043509
Log-likelihood        -1626,410   Akaike criterion     3256,821
Schwarz criterion      3265,378   Hannan-Quinn         3260,169
```

In the table, the OLS estimate of the linear regression model $\widehat{\texttt{wage}} = \hat{\alpha} + \hat{\beta}$ `experience` = 8.385 + 0.036$\cdot$`experience` is presented (the numbers $\hat{\alpha}\,(=8.385)$ and $\hat{\beta}\,(=0.036)$[4] are the *estimates* of unknown coefficients $\alpha$ and $\beta$). There are a few other important numbers here: the $p$-

[1] Clearly, when the `experience` (and age!) increases, the `wage` ultimately begins to decline.

[2] The red regression curve in Fig. 3.1 is unknown. We use our sample points to estimate it. The discrepancy between the estimated regression curve (not shown in the figure) and the observed value of `wage` is called a *residual* and denoted $\hat{\varepsilon}_i$. If our model is „good", the difference between the true and estimated regression curves (i.e., between errors $\varepsilon_i$ and residuals $\hat{\varepsilon}_i$) should not be big.

[3] To find the solution, take partial derivatives of $SSR$ with respect to $\hat{a}$ and $\hat{b}_1$ and equate them to zero (GRETL knows and uses relevant procedures).

[4] $\hat{\beta}$ =0.036 which means that if a worker's experience increases by 1 (year), his or her salary increases by 0.036 (dollars per hour).

value 0.0435 informs us that the hypothesis $H_0: \beta = 0$ must be rejected[5], i.e., the term `experience` is *significant*[6] in our model, i.e., we cannot remove `experience` from the model. The *coefficient of determination*[7] `R-squared` is always between 0 and 1 (the more the better), it indicates that `experience` explains only 0.765% of `wage` variation (this means that 99.235% of this variation remains unexplained[8] – it is the first indication that our model is not satisfactory, it lacks some important ingredients).

Our OLS procedure estimates the coefficients and $p$ - values correctly only if a model satisfies certain conditions. The most important are:

1) The spread (variance) of errors must be the same across observations
2) If the variables are time series, the errors must be uncorrelated, and
3) The errors must have a distribution close to the normal distribution.

Prior to testing these hypotheses, in order to get some intuition about the model, we shall plot a scatter plot with a regression line. In the Model 1 window, go to Graphs * Fitted, actual plot * Against experience:

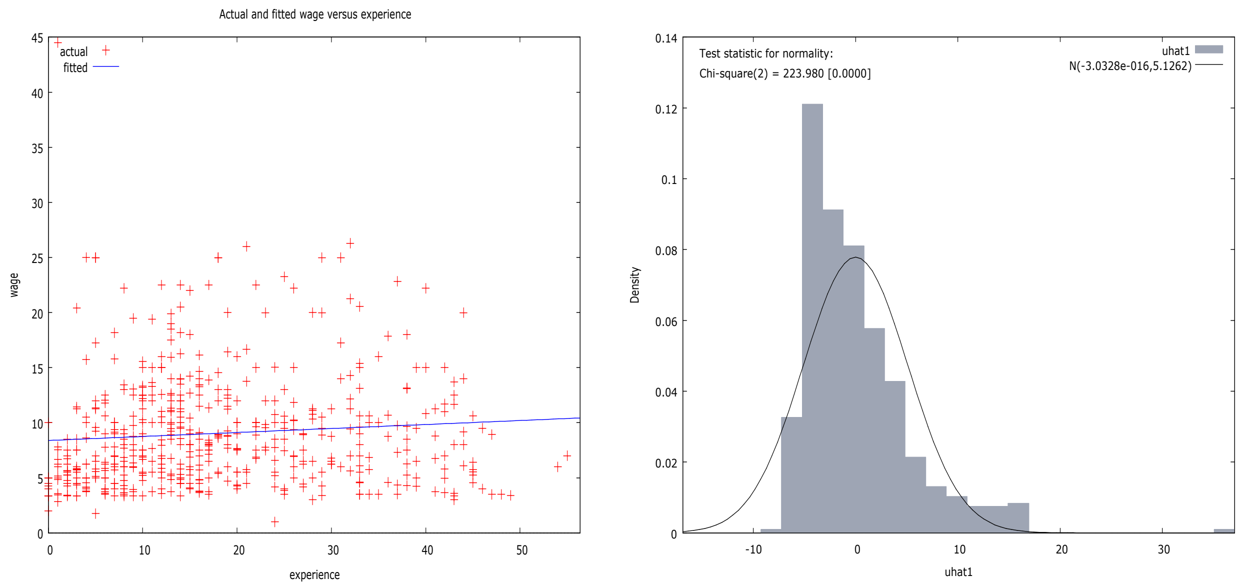



Fig. 3.2. `Experience-wage` scatter diagram and linear regression line (left). Naturally, when a worker gets older (that is, his `experience` increases), his wage diminishes (thus, economically speaking, a parabolic model is more appropriate.)

Figure 3.2 is rather informative. Firstly, the points above the blue regression line digress further upwards than those below the line downwards (this means that, probably, residuals are non-normal). Indeed, if you go, in the model window, to Tests| Normality of residual, you will get a histogram

---

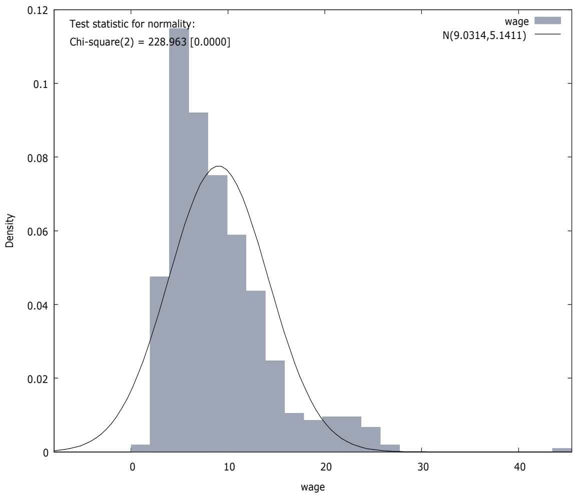
[5] Because it is less than 0.05.

[6] Definition. The variable $X$ is *significant* in the model $Y = \alpha + \beta X + \varepsilon$ if $\beta \neq 0$. To test this assumption, i.e., the hypothesis $H_0: \beta = 0$, see the fifth column in the regression printout: if the $p-$value is less than 0.05, $H_0$ is rejected, i.e., $X$ is significant.

[7] *R*-squared measures how accurately *Y* can be explained in terms of *X*'s.

[8] The factors which could explain these 99.235% of variation are hidden in $\varepsilon$.

shown in Fig. 3.2, right, with the $p$ - value of 0.0000 which means that we reject normality. In fact, this is a consequence of a highly skewed distribution of wage (see Fig. 3.3, left); usually we can correct the situation by creating a model for l_wage = log(wage) instead of for wage:



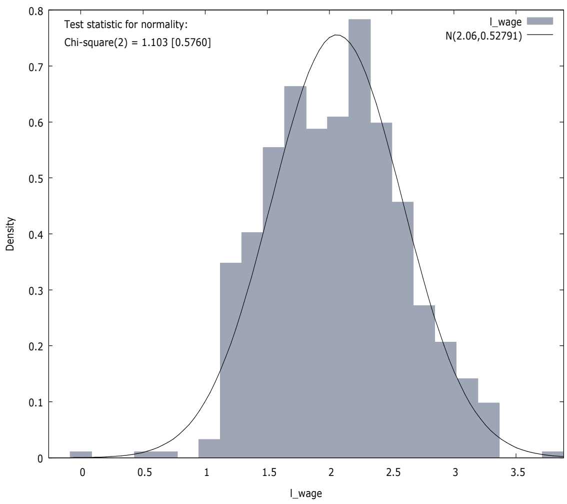




Fig. 3.3. Histogram of non-normal wage (left; there always are some rich and very rich people) and histogram of normal l_wage (right)

```
Model 2: OLS, using observations 1-533
Dependent variable: l_wage

               coefficient   std. error   t-ratio    p-value
  -----------------------------------------------------------
  const         1.97790       0.0398764    49.60     1.80e-201 ***
  experience    0.00460775    0.00183808    2.507    0.0125    **

R-squared            0.011696   Adjusted R-squared   0.009835
F(1, 531)            6.284160   P-value(F)           0.012480
Log-likelihood      -412.1606   Akaike criterion     828.3213
Schwarz criterion    836.8783   Hannan-Quinn         831.6698
```

The model has not improved much but its residuals are normal now (test it yourself). Also, the **interpretation of coefficients has changed**: in the linear-linear Model 1, namely, wage = 8.385 + 0.036 experience, if experience increases by 1 (year), wage increases 0.036 (dollars per hour) and in the log-linear Model 2: l_wage = 1.978 + 0.005 experience, if experience increases by 1, wage increases (0.005*100%=) 0.5%.

Secondly, the economic considerations suggest that when a person becomes elder, his/her salary begins to decrease, therefore we must either look for another functional form[9] of the dependence or include new variables into the model. The latter variant leads to multivariate regression and will be discussed later, now we replace linear depence by parabolic.

To create a quadratic model, we have to append the list of our variables with a square of experience: in GRETL's window select experience and go to Add * Squares of selected

[9] That is, for not a straight line (try, for example, a parabola.)

variables * OK (a new variable, sq_experience, will appear in GRETL's window.) Now go to Model * Ordinary least squares… * fill Dependent variable box with wage and append Independent variables box with sq_experience * OK.

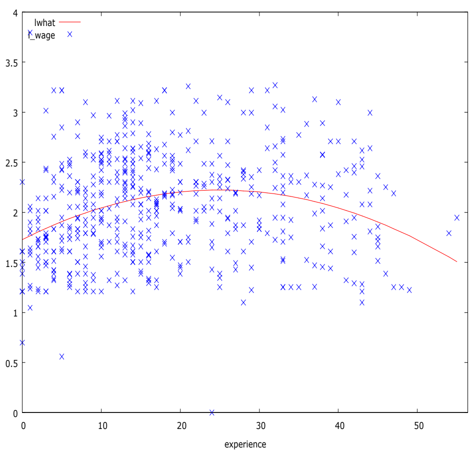
```
Model 3: OLS, using observations 1-533
Dependent variable: l_wage

                    coefficient    std. error    t-ratio    p-value
  ---------------------------------------------------------------
  const              1.72772       0.0575627     30.01     2.18e-116 ***
  experience         0.0395902     0.00622202     6.363    4.29e-010 ***
  sq_experience     -0.000792656   0.000135072   -5.868    7.76e-09  ***

Sum squared resid      137.5874   S.E. of regression   0.509508
R-squared              0.071996   Adjusted R-squared   0.068494
F(2, 530)              20.55906   P-value(F)           2.52e-09
Log-likelihood        -395.3834   Akaike criterion     796.7668
Schwarz criterion      809.6024   Hannan-Quinn         801.7896
```

Both variables (experience and sq_experience) in Model 3 are significant[10]. To choose between two or more models with the same left-hand side variable, use the Akaike and/or Schwarz criterions (choose the one with the smallest value of the criterion, thus, in our case select Model 3). Note that its R-squared is still very low, we will improve the model in the next section.

Model 3 is a model for $\log(\mathtt{wage})$. To get back to wage, we have to take antilogarithms; more specifically, $\widehat{\mathtt{wage}} = \exp(\widehat{\log}(\mathtt{wage}) + \hat{\sigma}_\varepsilon^2 / 2)$, that is perform the following commands (copy and paste the following text to the GRETL script window):

```
ols l_wage const experience sq_experience
lwhat = $yhat                     # fitted values in Model 3
what = exp($yhat+$sigma^2/2)      # back to wage
```

(can you plot the two graphs shown below?)



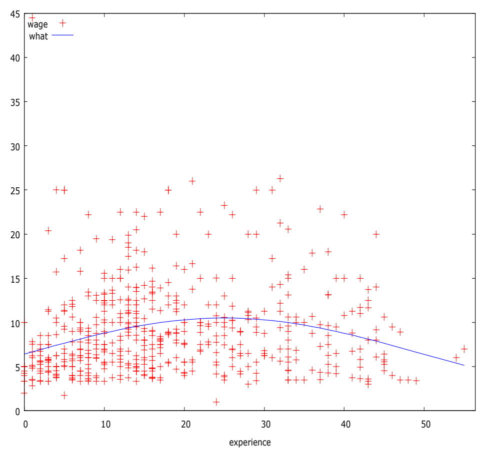




Fig. 3.4. Two parabolic regression curves – in experience-log(wage) scale (left) and experience-wage scale (right)

[10] Note the rule – if the squared term is significant, never remove linear term.

**3.1 exercise.** Regress experience on age and analyse the model. How do you interpret the coefficient $\hat{\beta}$ ? Add sq_age to the model and analyze it. Which of the two models is better? (to choose, use Akaike's criterion and also analize normality of residuals).

**3.2 exercise.** Import the file houseprice.xls where

```
price          price of a house in dollars
assess         assessed value in dollars
bdrms          number of bedrooms
lotsize        size of lot - square feet
sqrft          size of house - square feet
colonial       =1 if home is colonial style
lprice         log(price)
lassess        log(assess)
llotsize       log(lotsize)
lsqrft         log(sqrft)
```

Regress price on lotsize. Analyze the model. Draw relevant graphs. You can see than one observation corresponding to a lotsize above 90000 distorts our model. To remove the observation go to Data * Sort data... * Select sort key „lotsize" * OK; now the abnormal observation is the last, namely, 88th; go to Sample * Set range... * End: „87" * OK. Again regress price on lotsize. Compare the two models. Which one is better? Create one more model – regress price on sqrft. Analyze it. ◄◄

Once we have created a regression model, it is easy to use it to predict the response variable. For example, in order to employ Model 3: wage=exp(1.728+0.040*experience-0.001*experience^2+0.510^2/2), assume that experience of a new worker is 35 years[11]; it is easy to calculate that predicted wage equals 7.637. Later, provided we know more about the worker, we shall be able to improve our prediction.

*******************************************
*******************************************

Recall now that, in the equation $\hat{y} = \hat{\alpha} + \hat{\beta}x$, the coefficient $\hat{\beta}$ tells you by how many units y is expected to increase when x increases by one unit. **A special attention to this interpretation should be drawn in the case where x is a nominal variable.** Namely, is it true that men (gender=2) and women (gender=1) get different wages for the same work? To test this claim, we have first to convert gender to two dummy variables: select gender and go to Add * Dummies for selected discrete[12] variables * OK. Two new dummy[13] variables will be created: Dgender_1 equals 1 for women and 0 for men, and Dgender_2 which equals 1 for men and 0 for women. Next, to create a regression model, go to Model etc and create the model[14] $wage = \alpha + \beta * Dgender\_2 + \varepsilon$; its estimate is given below:

[11] Avoid using the value of explanatory variable outside its range; in our case, experience is between 0 and 55.
[12] If the variable x is to denote *k* groups and thus takes any *k* numeric values, we should tell gretl that these values are not numbers. In order to do this, sometimes you have to select x, go to Variable| Edit attributes and check the „Tick this variable as discrete" box.
[13] The *dummy* (or indicator) *variable* is to designed to mark the group of cases where the variable equals 1.
[14] In a model with dummy variables, one dummy variable should always be excluded (our model does not contain Dgender_1.)

```
Model 3: OLS, using observations 1-533
Dependent variable: wage

              coefficient   std. error   t-ratio    p-value
  -----------------------------------------------------------
  const        7,89025      0,322497     24,47     4,35e-089 ***
  Dgender_2    2,10467      0,437966      4,806    2,01e-06  ***

Mean dependent var   9,031426   S.D. dependent var   5,141105
Sum squared resid    13475,23   S.E. of regression   5,037567
R-squared            0,041678   Adjusted R-squared   0,039873
F(1, 531)            23,09329   P-value(F)           2,01e-06
Log-likelihood      -1617,112   Akaike criterion     3238,225
Schwarz criterion    3246,782   Hannan-Quinn         3241,573
```

Thus, $\widehat{\texttt{wage}} = \texttt{7.890+2.105*Dgender\_2}$; it means that the mean wage in the base group Dgender_2=0 (that is, the women group) equals 7.890 and in the men's group (Dgender_2=1) it is $2.105/hour bigger. What is more important, this increase is significant, which proves that we cannot reject[15] the above claim. Now, to get a deeper insight, sort your data by Dgender_2, redo the model and plot a graph via Graphs * Fitted,actual plot * By observation number – the graph is not very convincing in proving that men's salary is bigger; this means that we have to distinguish the concept of "statistically different" and "economically different".

**3.3 exercise.** The same conclusion we can get with the $t$ - test (test the null that men's wage is the same as that of women.)

**3.4 exercise.** Is it true that white people (or caucasians or ethnicity=1) earn most for the same job? (Recode ethnicity to three dummy variables and create a regression model for wage with only <u>two</u>[16] dummies for hispanic and other.) Are the coefficients[17] at these dummies negative? Are they significant? Plot some graphs. Comment your findings.

**3.5 exercise.** Use the ANOVA procedure to test the null that wage is the same in the three ethnic groups[18]. Is the the hypothesis accepted? ◀◀

## 3.2. Multiple Regression

Simple or univariate linear regression was used to model the effect one variable, an explanatory variable, has on another, the response, variable. In particular, if an explanatory variable changes by some amount, the response changes by a multiple of that same amount (that multiple being the slope of the regression line.) Multiple linear regression does the same, only there are multiple explanatory variables.
There are many situations where intuitively this is the correct model. For example, the price of a new hause depends on many factors (the number of bedrooms, the number of bathrooms, the location of the house, etc.) When a house is built, it costs a certain amount for the builder to build an extra room and so the cost of house reflects this. In fact, in some new developments, there is a pri-

[15] We shall be more specific about this hypothesis in the next section.
[16] If your variable tranforms to $k$ dummy variables, the model must contain only $k$ - 1 dummy variables.
[17] These coefficients mean extra payment (compared with caucasian) for being hispanic or other.
[18] Regression models not only detect the existance of differences in means but also gives the values of these differences and their significance.

celist for extra features such as $900 for an upgraded fireplace. Now, if you are buying an older house it isn't so clear what the price should be. However, people do develop rules of thumb to help figure out the value. For example, these may be add $30,000 for an extra bedroom and $15,000 for an extra bathroom, or subtract $10,000 for the busy street. These are intuitive uses of a linear model to explain the cost of a house based on several variables.

So far we have investigated models of the form `wage`$=f($`experience`$)+\varepsilon$. After a while, we shall return to the multiple model `wage`$=f($`education,experience,age,...`$)+\varepsilon$ or, in other words, to the model where the right-hand-side contains many explanatory variables.

The basic ($k$ - variate) model is of the form $y = f(\mathbf{x}) = \alpha + \beta_1 x_1 + ... + \beta_k x_k + \varepsilon$[19]. Note that there are many variants of this model with $x_1$ replaced by, for example, $\log(x_1)$ or with the right-hand-side containing $x_1^2$ or $x_1 * x_3$ etc. We begin with a rather simple multivariate regression model.

**3.1 example.** Import once again houseprice.xls:

| | |
|---|---|
| `price` | price of a house in dollars |
| `assess` | assessed value in dollars |
| `bdrms` | number of bedrooms |
| `lotsize` | size of lot - square feet |
| `sqrft` | size of house - square feet |
| `colonial` | =1 if home is colonial style, 0 otherwise |

A common approach to begin is to include all the variables into the newly created multiple regression model. Thus, to create the model $price = \alpha + \beta_1 \cdot bdrs + \beta_2 \cdot lotsize + \beta_3 \cdot sqrft + \beta_4 \cdot colonial + \varepsilon$, go to Model * Ordinary Least Squares..., choose price as Dependent variable and the rest as Independent:

```
Model 1: OLS, using observations 1-88
Dependent variable: price

             coefficient    std. error    t-ratio    p-value
  ----------------------------------------------------------
  const      -24126.5       29603.5       -0.8150    0.4174
  bdrms       11004.3        9515.26       1.156     0.2508
  lotsize         2.07583       0.642651   3.230     0.0018    ***
  sqrft         124.237        13.3383     9.314     1.53e-014 ***
  colonial    13715.5       14637.3        0.9370    0.3515

Sum squared resid    2.98e+11   S.E. of regression   59876.97
R-squared            0.675792   Adjusted R-squared   0.660167
F(4, 83)             43.25210   P-value(F)           1.45e-19
Log-likelihood      -1090.297   Akaike criterion     2190.594
Schwarz criterion    2202.981   Hannan-Quinn         2195.584

Excluding the constant, p-value was highest for variable 6 (colonial)
```

[19] The meaning of the coefficient $\beta_i$ is now slightly different from that of a univariate case: it shows by how many units $y$ changes if $x_i$ increases by 1 unit, ceteris paribus (these latin words mean “provided all other variables do not change”). Thus, $\beta_i$ shows the isolated effect of $x_i$ on $y$.

As you can see, the least significant term is `colonial` – exclude it from the model (in the Model 1 window go to Tests * Omit variables).

```
Model 2: OLS, using observations 1-88
Dependent variable: price

             coefficient    std. error    t-ratio    p-value
  ----------------------------------------------------------
  const       -21770.3       29475.0       -0.7386    0.4622
  bdrms        13852.5        9010.15       1.537     0.1279
  lotsize          2.06771       0.642126   3.220     0.0018    ***
  sqrft          122.778        13.2374     9.275     1.66e-014 ***

Mean dependent var   293546.0   S.D. dependent var   102713.4
Sum squared resid    3.01e+11   S.E. of regression   59833.48
R-squared            0.672362   Adjusted R-squared   0.660661
F(3, 84)             57.46023   P-value(F)           2.70e-20
Log-likelihood      -1090.760   Akaike criterion     2189.520
Schwarz criterion    2199.429   Hannan-Quinn         2193.512

Excluding the constant, p-value was highest for variable 3 (bdrms)

Comparison of Model 1 and Model 2:

  Null hypothesis: the regression parameter is zero for colonial
  Test statistic: F(1, 83) = 0.878023, with p-value = 0.351462
  Of the 3 model selection statistics, 3 have improved.
```

Next, we shall exclude `bedrooms`:

```
Model 3: OLS, using observations 1-88
Dependent variable: price

             coefficient   std. error    t-ratio    p-value
  ---------------------------------------------------------
  const       5932.41       23512.4        0.2523    0.8014
  lotsize        2.11349        0.646560   3.269     0.0016    ***
  sqrft        133.362         11.3969    11.70      2.11e-019 ***

Mean dependent var   293546.0   S.D. dependent var   102713.4
Sum squared resid    3.09e+11   S.E. of regression   60311.54
R-squared            0.663143   Adjusted R-squared   0.655217
F(2, 85)             83.66618   P-value(F)           8.25e-21
Log-likelihood      -1091.981   Akaike criterion     2189.962
Schwarz criterion    2197.394   Hannan-Quinn         2192.956

Comparison of Model 2 and Model 3:

  Null hypothesis: the regression parameter is zero for bdrms
  Test statistic: F(1, 84) = 2.36371, with p-value = 0.127945
  Of the 3 model selection statistics, 2 have improved.
```

> If you have several models to explain $Y$, choose the one with all significant terms and minimal Schwarz and/or Akaike[20] criterion value

[20] Both criteria are based on RSS (that is, `Sum squared resid`), hence, the smaller the better. On the other hand, these criteria also add some penalty on the number of variables in the model, therefore we can compare models with different right hand sides.

According to our rule, we choose model 3. Note the meaning of the coefficient at `lotsize`: if the `lotsize` increases 1 (square foot), the price will increase 2.11349 (dollars) provided all the other variables (in our case it is only `sqrft`) do not change. Similarly, for each change of 1 sqft in `sqrft`, the `price` increases 133.362 dollars (if sizes of the houses remain the same).

*****************

One of the possible applications of the regression models is forecasting. What price our model predicts for a house whose lotsize is 10000 and sqrft 3000? We can calculate the price manually:

$$price = 5932.41 + 2.11349 * 10000 + 133.362 * 3000 = 427153.3$$

On the other hand, GRETL can also help you: go to Data * Add observations * 1; then click on `lotsize`, choose Edit Values..., insert 10000, press Enter, click on `sqrft`, choose Edit Values..., insert 3000, go to Model 3, choose Analysis * Forecasts... * OK:

```
For 95% confidence intervals, t(85, 0.025) = 1.988

     Obs          price    prediction    std. error          95% interval

      46      265000.00     257740.13
      ..................................................................
      88      242000.00     252978.43
      89                    427153.41     61668.114     304540.68 - 549766.15
```

thus, we get the same prediction.

**3.6 exercise**. Import once again CPS1985.txt. Start with a model where all the variables will serve as explanatory for `wage`. Note that, for example, `ocupation` is a nominal variable and we have to convert it to a set of dummy variables (use `dummify` function to perform this task).

Below is an example of GRETL script (program). Copy these lines to command script window (open it with a click on the second from the left icon, see Fig. 1.1) and then click the pinion icon in the top line to run the script.

```
series exp=experience
series exp2=exp*exp
ols wage 0 education exp exp2 age dummify(ethnicity) dummify(region)
dummify(gender) dummify(occupation) dummify(sector) dummify(union)
dummify(married)
```

After some experimenting, I improved this model and ended with

```
ols wage 0 education exp exp2 dummify(region) dummify(gender)
dummify(occupation) dummify(union)

series wh2 = $yhat # wh2 is the predicted wage values
```

Compare prediction with the true values: draw a scatter diagram of `wage` (true values) and `wh2` (predicted values). If the model gives ideal prediction, the points will be on a diagonal line. What about our case?

Can you create a still better model[21]?

**3.7 exercise**. In exam.txt you will find the scores in the final examination $F$ and the scores in two preliminary examinations $P_1$ and $P_2$ for 22 students in a statistics course.

(a) Fit each of the following models to the data:

Model 1: $F = \alpha + \beta_1 P_1 + \varepsilon$

Model 2: $F = \alpha + \beta_1 P_2 + \varepsilon$

Model 3: $F = \alpha + \beta_1 P_1 + \beta_2 P_2 + \varepsilon$

(b) Which of the three models is the best?
(c) Use the best model to predict the final examination score for a student who scored 78 and 85 on the first and second preliminary examinations, respectively.

**3.8 exercise**. A national organization wanted to study the consumption pattern of cigarettes in all 50 U.S. states and the District of Columbia. The data is given in the file CigCons.txt, it contains the following variables:

| | |
|---|---|
| `Age` | median age of person living in a state |
| `HS` | percentage of people over 25 years of age in a state who had completed high school |
| `Income` | per capita personal income for a state |
| `Black` | percentage of blacks living in a state |
| `Female` | percentage of females living in a state |
| `Price` | weighted average price of a pack of cigarettes in a state |
| `Sales` | number of packs of cigarettes sold in a state on a per capita basis |

(a) Is the variable `HS` needed in the regression equation relating `Sales` to the six predictor variables?
(b) Improve the model by removing insignificant variables. What is your final model?
(c) What is the meaning of the coefficient at `Price` in this model?

**3.9 exercise**. In a statistics course, personal information was collected on all the students for class analysis. Data on `Age` (in years), `Height` (in inches), and `Weight` (in pounds) of the students are given in WvsH.txt. The sex of each student is also noted as `Female` and coded as 1 for women and 0 for men. We want to study the relationship between the height and weight of students (weight is the response variable and height predictor variable).

(a) Draw a scatter diagram of `Weight` vs `Height` (to mark women and men differently, go to View| Graph specified vars| X-Y with factor separation… and use Female as Factor). Comment the scatter diagram.
(b) Create the following two equations:

Model 1: $Weight = \alpha + \beta_1 Height + \varepsilon$

Model 2: $Weight = (\alpha + \gamma_1 Female) + \beta_1 Height + \varepsilon$

[21] Better means a model with all significant terms and still smaller value of Akaike's and/or Schwarz's criteria.

(Model 2 is, in fact, described by two regression lines – the intercept of the first, for men, equals $\alpha$ and the second, for women, equals $\alpha+\gamma_1$; slope of the regression lines in both cases is the same, $\beta_1$).

(c) Which of the two models is better? Why? What is the meaning of $\gamma_1$ in Model 2? Draw a scatterplot of `Weight` vs `Height` with two regression lines, one for women and one for men (go to Graphs| Fitted,actual plot| Against `Height` in the regression model window).

(d) Create a new variable `FH=Female*Height` and analyse the

Model 3: $Weight = (\alpha+\gamma_1 Female)+(\beta_1+\gamma_2 Female)\cdot Height+\varepsilon =$
$$(\alpha+\gamma_1 Female)+\beta_1 Height+\gamma_2 FH+\varepsilon$$

Now the sex (probably) affects not only the intercept, but also the slope of the regression line. What can you tell about Model 3? What is your final model? ◄◄

We have already mentioned (see p. 3-4) that in order to get correct estimates of the coefficients and their $p$- values a model must satisfy certain conditions, specifically, the errors must be close to normal. The most common deflection from normality is the right skewness of residuals (this means that the tail on the right side of the histogram is longer or fatter than the left side; in other words, there are too many too big values in our sample[22]). Transforming the outcome is often successful for reducing the skewness of residuals. The rationale is that the more extreme values of the outcome are usually the ones with large residuals (defined as $\hat{\varepsilon}_i = y_i - \hat{y}_i$); if we can "pull in" the outcome values in the tail of the distribution toward the center, then the corresponding residuals are likely to be smaller too (one such transformation is to replace the outcome $y$ with $\log y$). Recall that if we fit the linear-linear regression model $y=\alpha+\beta x+\varepsilon$ or, what is the same, $E(y\,|\,x)=\alpha+\beta x$, the increase of one-unit in $x$ is associated with $\beta$ units change in $y$. If we include logarithms into the model, we use the following terminology (note that logarithm always means percentage):

2. Linear-log model is $\hat{y}=\hat{\alpha}+\hat{\beta}\log x$. We would say that a one <u>percent</u> change in $x$ leads to an (approximately) $\hat{\beta}/100$ <u>unit</u> change in $y$;

3. Log-linear model is $\widehat{\log(y)}=\hat{\alpha}+\hat{\beta}x$ – a one <u>unit</u> change in $x$ leads to an (approximately) $100\hat{\beta}$ <u>percentage</u> change in $y$;

4. Log-log model is $\widehat{\log(y)}=\hat{\alpha}+\hat{\beta}\log(x)$ – a one <u>percent</u> change in $x$ leads to an (approximately) $\beta$ <u>percentage</u> change in $y$.

**3.2 example.** Let us consider the data set CPS1985.txt once again. In the model's

```
Model 1: OLS, using observations 1-533
Dependent variable: wage

              coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const        -0.742830    1.05070      -0.7070    0.4799
  education     0.750242    0.0790819     9.487     7.95e-020 ***
```

[22] If our variable is `wage`, there are often some millionaires in the sample.

window (what is the meaning of 0.75?), go to Tests| Normality of residual – the $p$ - value of normality test is 0.0000<0.05, thus we reject the normality hypothesis. Next, add a new variable `l_wage` and create new model

```
Model 2: OLS, using observations 1-533
Dependent variable: l_wage

               coefficient   std. error   t-ratio    p-value
  ----------------------------------------------------------
  const        1.06078       0.107971      9.825     4.82e-021 ***
  education    0.0766965     0.00812651    9.438     1.19e-019 ***
```

Now the $p$ - value is 0.89, thus we could expect that Model 2 is more accurate (what is the meaning of the coefficient 0.077?). ◄◄

**3.11 exercise**. The Western Collaborative Group Study (WCGS), a prospective cohort study, http://clinicaltrials.gov/show/NCT00005174, recruited middle-aged men (ages 39 to 59) who were employees of 10 California companies and collected data on 3154 individuals during the years 1960-1961 (see). These subjects were primarily selected to study the relationship between the "type A" behavior pattern and the risk of coronary hearth disease (CHD). A number of other risk factors were also measured to provide the best possible assessment of the CHD risk associated with behavior type. Additional variables collected include age, height, weight, systolic blood pressure, diastolic blood pressure, cholesterol, smoking, and corneal arcus. The data can be found in the wcgs.txt file.

| | |
|---|---|
| `no` | number of observation |
| `chd69` | had a coronary heart desease (CHD) event (Yes or No) |
| `chd` | =1 if `chd69`=Yes and =0 if `chd69`=No |
| `arcus` | presence/absence of corneal arcus senilis (`1` or `0`) |
| `behpat` | self-reported behavior pattern (A1, A2, B3, B4) (risk factor for CHD) |
| `bmi` | body mass index (weight in kg divided by the square of height in meters) |
| `chol` | baseline LDL (low-density lipoprotein) cholesterol levels (mg/100ml) |
| `dbp` | diastolic blood pressure |
| `dibpat` | Dichotomous behavior pattern: 0 = Type B; 1 = Type A |
| `height` | height of a patient (inches) |
| `lnsbp` | logarithm of `sbp` |
| `lnwght` | logarithm of `weight` |
| `ncigs` | cigarettes per day |
| `sbp` | systolic blood pressure |
| `smoke` | smokes or no (`Yes` or `No`) |
| `typchd69` | type of CHD event (0, 1, 2, 3) |
| `age` | age |
| `agec` | age categories 35-40, 41-45, 51-55, 56-60 |
| `age_cat` | `agec` coded, respectively, as 1, 2, 3, 4 |
| `weight` | weight of a patient (lbs) |
| `wghtcat` | weight categories <140, 140-170, 170-200, >200 |
| `wg_cat` | `wghtcat` coded, respectively, as 1, 2, 3, 4 |

1) Test `dbp` and `l_dbp` for normality (use frequency distribution, boxplot, Variable| Normality test, Variable| Normal Q-Q plot…).
2) Estimate the frequency distribution of the nominal variable `behpat`. Can we test for normality in this case?
3) Use any relevant procedure to test whether `behpat` influences `sbp`.
4) Does `weight` influences `sbp`?
5) With categorical variables, the typical method is to tabulate the outcomes. Estimate the cross-tabulation table of `behpat` and `wghtcat`. Are these two variables associated?
6) Is `sbp` the same in each `wghtcat` group?
7) How many patients smoke?
8) Is there any relationship between `behpat` and `chd69`? And `smoke` and `chd69`? ◄◄

## 3.3. Logit Regression

Recall that in simple linear regression, we modeled the average of a continuous outcome variable as a function of a single continuous or discrete predictor, using a linear relationship of the form $\texttt{wage} = \alpha + \beta\,\texttt{experience} + \varepsilon$ which can also be written as[23] $E(\texttt{wage}|\texttt{experience}) = \alpha + \beta\,\texttt{experience}$ (we used the OLS method to estimate $\alpha$ and $\beta$). Now, consider the regression model $\texttt{chd} = \alpha + \beta\,\texttt{age} + \varepsilon$ (see 3.11 exercise) where, in contrast to the previous case, the response variable `chd` takes only two values, 1 and 0. As it follows from the probability theory, $E(\texttt{chd}) = P(\texttt{chd} = 1)$, therefore the regression line $(E(\texttt{chd}\,|\,\texttt{age}) =)$ $P(\texttt{chd=1}|\texttt{age}) = \alpha + \beta\,\texttt{age}$ obtained with the usual OLS procedure, describes how the risk (or probability) of the coronary heart desease depends on `age`.

```
Model 1: OLS, using observations 1-3154
Dependent variable: chd

             coefficient   std. error    t-ratio    p-value
  ---------------------------------------------------------
  const      -0.191903     0.0408262     -4.700     2.71e-06   ***
  age         0.00590741   0.000875965    6.744     1.83e-011  ***

R-squared            0.014224   Adjusted R-squared   0.013911
Log-likelihood     -364.6099    Akaike criterion     733.2199
```

This, the so-called *linear probability model*, has some drawbacks – for example, if one wants to fit the probability $P(\texttt{chd} = 1\,|\,\texttt{age} = 30)$, he or she will get a negative value (see Fig. 3.5, left) what is impossible for probability.

[23] And read as the ”expectation of `wage` for the given value of `experience`“.

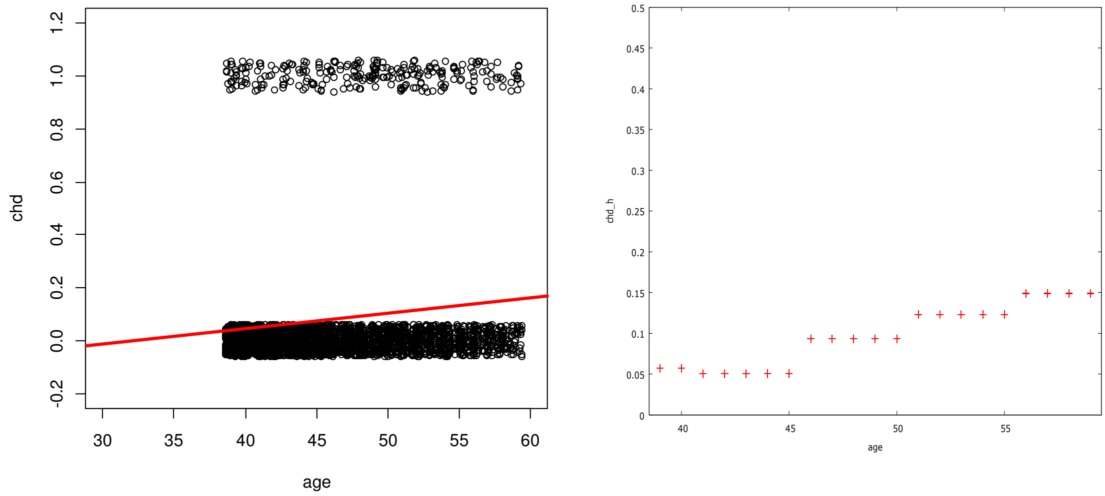



Fig. 3.5. Linear probability model (left; to enhance readabily of the graph, we jittered the data points); probability model for grouped data (right, "+" marks the estimate of the risk probability in each group of `age_cat` (the probability of the disease in the 35-40 group is a bit strange)

We can avoid this trap by grouping our data by `age` (this is what the variables `agec` or `age_cat` are meant for). To estimate the probability of heart desease event in each `age_cat` group is easy – one has to calculate the relative frequency of 'Yes' in respective group.

To do this in GRETL, act as follows: mark `chd` as discrete (right-click on it and choose Edit attributes), dummify it, and create respective regression model. In Hansl, the GRETL scripting language, we use the following script (copy and paste the text into the cripting window):

```
ols chd 0 dummify(age_cat)
series chd_h = $yhat                # fitted probabilities in groups
gnuplot chd_h age --output=display --suppress-fitted
```

```
Model 2: OLS, using observations 1-3154
Dependent variable: chd

               coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
  const          0.0570902   0.0116624     4.895    1.03e-06 ***
  Dage_cat_2    -0.00667777  0.0142725    -0.4679   0.6399
  Dage_cat_3     0.0362431   0.0153129     2.367    0.0180   **
  Dage_cat_4     0.0660158   0.0166098     3.975    7.21e-05 ***
  Dage_cat_5     0.0916701   0.0210046     4.364    1.32e-05 ***

Mean dependent var   0.081484   S.D. dependent var   0.273620
Sum squared resid    232.5669   S.E. of regression   0.271761
R-squared            0.014792   Adjusted R-squared   0.013540
F(4, 3149)           11.81970   P-value(F)           1.57e-09
Log-likelihood      -363.7008   Akaike criterion     737.4017
Schwarz criterion    767.6838   Hannan-Quinn         748.2664
```

Recall the meaning[24] of these coefficients:

$$(P(\texttt{chd69}='Yes'|\texttt{age\_cat})=)\ \underline{P(\texttt{chd}=1|\texttt{age\_cat})}\ \ (=E(\texttt{chd}|\texttt{age\_cat}))=$$

$$\begin{cases} 0.057 & \text{if } \texttt{age\_cat}=1 \\ 0.057-0.007 & \text{if } \texttt{age\_cat}=2 \\ 0.057+0.036 & \text{if } \texttt{age\_cat}=3 \\ 0.057+0.066 & \text{if } \texttt{age\_cat}=4 \\ 0.057+0.092 & \text{if } \texttt{age\_cat}=5 \end{cases}$$

The plot of this model (see Fig. 3.5, right) gives a better understanding of how categorized age affects the probability of the heart disease. However, note that we grouped (rounded) the values of age which means that we lost some information contained in it. To use the information on age in a more precise manner, we have to take into account the exact year of the age. Unfortunately, if we narrow the grouping intervals (and finally stop at every single year), the accuracy of the estimated probability in each group will deteriorate (there will be too many parameters, i.e., probabilities, to estimate). Therefore we now take still another approach to estimate these probabilities: instead of a straight line, we want to use any monotone curve taking values between 0 and 1. The most popular curve is the so-called logistic curve $y=\Lambda(x)=\exp(x)/(\exp(x)+1)$, $-\infty<x<\infty$, (see its graph in Fig.3.6, left). More specifically, we want to approximate our age-chd data with the curve $P(\texttt{chd=1}|\texttt{age})=\Lambda(\alpha+\beta\,\texttt{age})$ where the parameters $\alpha$ and $\beta$ are still to be estimated. To do this,

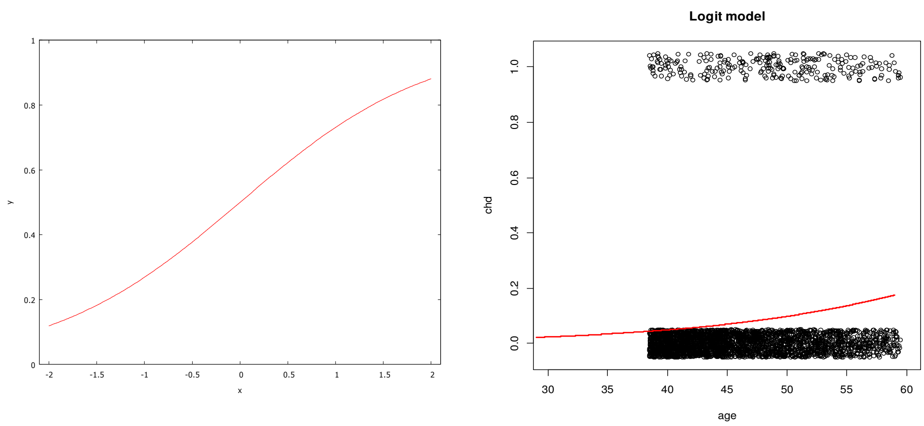



Fig. 3.6. Logistic curve (left) and the age-chd logit Model 3 (right)

we have to use another[25] method (not the OLS) but, in any case, GRETL knows how to do this –go to Model| Limited dependent variable| Logit| Binary:

```
Model 3: Logit, using observations 1-3154
Dependent variable: chd
```

[24] Note that now we had to estimate five unknown parameters (instead of two in previous model).
[25] It is the so-called maximum likelihood method.

```
              coefficient   std. error      z         slope
  -----------------------------------------------------------
  const       -5.93952      0.549323     -10.81
  age          0.0744226    0.0113024      6.585    0.00523797

Mean dependent var   0.081484   S.D. dependent var   0.273620
McFadden R-squared   0.024077   Adjusted R-squared   0.021832
Log-likelihood      -869.1781   Akaike criterion     1742.356
Schwarz criterion    1754.469   Hannan-Quinn         1746.702


Number of cases 'correctly predicted' = 2897 (91.9%)
f(beta'x) at mean of independent vars = 0.070

           Predicted
               0      1
  Actual 0  2897      0
         1   257      0
```

(the plot produced by GRETL differs from the one given in Fig. 3.6, right, but it is essentially the same).

A few words about the model's printout. To estimate the accuracy of the model, on can use (McFadden's) R-squared (it is between 0 and 1, the more the better). The usual considerations for Akaike and Schwarz criteria also hold. Another goodness-of-fit measure is the percent 'correctly predicted': define a binary predictor of chd , that is $\widehat{\texttt{chd}}$ , to be one (we predict, a patient will suffer CHD event) if the predicted probability is at least 0.5 and zero otherwise. There are four possible outcomes on each pair $(\texttt{chd}_i, \widehat{\texttt{chd}_i})$ and when both are zeros or both are ones, we say that we made a correct prediction. The "Number of cases 'correctly predicted'" is the percentage of times that $\texttt{chd}_i = \widehat{\texttt{chd}_i}$ (in our example, it is 2897 (91.9%)). It may seem impressive, but here it just coincides with the number of chd=0 cases (in our case, the logistic curve never exceeds 0.5, thus we always predict $\widehat{\texttt{chd}_i} = 0$ ). To obtain a more sensible characteristic of the model, it is recommended, instead of 0.5, to use another threshold, namely, the percentage of zeroes which equals 0.081 (=8.1%). Now, the HANSL script

```
logit chd const age
series pr_chd=$yhat>0.081          # equals 0 or 1
scalar corr_pred=sum(chd-pr_chd=0)  # number of correct predictions
```

gives us a more useful "Number of cases 'correctly predicted'" which equals 1998 (63.3%).

Recall that in the linear probability model $P(\texttt{chd=1|age})$ = -0.192 + 0.006 age the *slope* coefficient $\hat{\beta}$ (=0.006 $\equiv d\,P(\texttt{chd}=1|\texttt{age})/d\,\texttt{age}$ ) showed the increment in the probability of $P(\texttt{chd=1|age})$ when age increased by 1 year. In the logit model, the slope is again defined as $d\,P(\texttt{chd}=1|\texttt{age})/d\,\texttt{age}$, but now this derivative depends also on age (differentiate $\Lambda(\alpha + \beta\,\texttt{age})$ with respect to age). In interpreting the estimated model, it is useful to calculate this value at, say, the sample mean of the independent variables (in our case, $d\,P(\texttt{chd}=1|\texttt{age})/d\,\texttt{age}$ $\approx$0.0744226*0.070=0.005).

The quality of prediction can be improved if we include more variables, for example, sbp and smoke (or rather, its dummified variant) to the rhs:

```
logit chd const age sbp dummify(smoke)
series pr_chd2=$yhat>0.081
scalar corr_pred2=sum(chd-pr_chd2=0)

Model 4: Logit, using observations 1-3154
Dependent variable: chd
Standard errors based on Hessian

             coefficient   std. error      z         slope
  -----------------------------------------------------------
  const      -8.31312      0.702393     -11.84
  age         0.0647724    0.0116839      5.544    0.00417883
  sbp         0.0238292    0.00377583     6.311    0.00153736
  Dsmoke_2   -0.662209     0.136071      -4.867   -0.0437844

Mean dependent var   0.081484   S.D. dependent var   0.273620
McFadden R-squared   0.057913   Adjusted R-squared   0.053422
Log-likelihood      -839.0431   Akaike criterion     1686.086
Schwarz criterion    1710.312   Hannan-Quinn         1694.778

Number of cases 'correctly predicted' = 2897 (91.9%)
f(beta'x) at mean of independent vars = 0.065
```

(now the scalar `corr_pred2` equals 2016; all the goodness-of-fit parameters have also improved compared with the previous model).

**3.12 exercise**. Go to File| Open data| Sample file…| Greene| greene19-1. This data set contains four variables:

```
  GPA          TUCE          PSI         GRADE
 2.66            20            0             0
 2.89            22            0             0
 3.28            24            0             0
 2.92            12            0             0
 4.00            21            0             1
.................................................................
```

where the data is taken from the study which examined whether a new method of teaching economics significantly influenced performance in later economics courses. Here

| | |
|---|---|
| `GPA` | student's grade point average |
| `TUCE` | the score on a pretest that indicates entering knowledge of the material test score on economics test |
| `PSI` | the binary variable indicator of whether the student was exposed to the new teaching method |
| `GRADE` | indicates the whether a student's grade in an intermediate macroeconomics course was higher (=1) than that in the principles course (dependent variable) |

Create a logit model explaining `GRADE` in terms of `GPA`, `TUCE`, and `PSI`. What is the number of cases 'correctly predicted' with a threshold equal to a) 0.5, b) the share of zeroes of `GRADE`? What is the increment $P(\texttt{GRADE}=1|\texttt{PSI}=1,...)-P(\texttt{GRADE}=1|\texttt{PSI}=0,...)$ at mean of independent variables? Draw two logistic curves as a function of `GPA` on a single graph: the first is for `PSI`=0 and the second for `PSI`=1 (both at the mean of `TUCE`). Comment your findings. ◄◄

# 4. Time Series Analysis

Most of the data we have analysed so far were the so-called **cross-sectional** data which were characterized by individual units and collected at more or less the same time. These units might refer to companies, people or countries. Another type of data are collected at specific points in time. In these examples, the data are ordered by time and are referred to as **time series** data. The underlying phenomenon which we are measuring (e.g., GDP, stock prices, interest rates, etc.) is referred to as a variable. Time series data can be observed at many frequencies. Commonly used frequencies are: annual (i.e., a variable is observed every year), quarterly (i.e., four times a year), monthly, weekly or daily. In contrast to cross-sectional data, tomorrow's value of time series correlates with past values which allows us to forecast the variable of interest. In this chapter, we shall analyse how one can achieve this goal.

## 4.1. Time Series: Examples

### 4.1 example. Stock returns.

Let $P_t$ be the price of an asset at time $t$. The one-period (simple[1]) *return* is the percentage change in price:

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}} * 100 .$$

Consider monthly returns on Citigroup stock from 1990:01 through 1998:12 (to input the data, open gretl, go to File * Open data * Sample file... * Ramanathan * data9-13, right-click on `cret` and choose Time series plot).

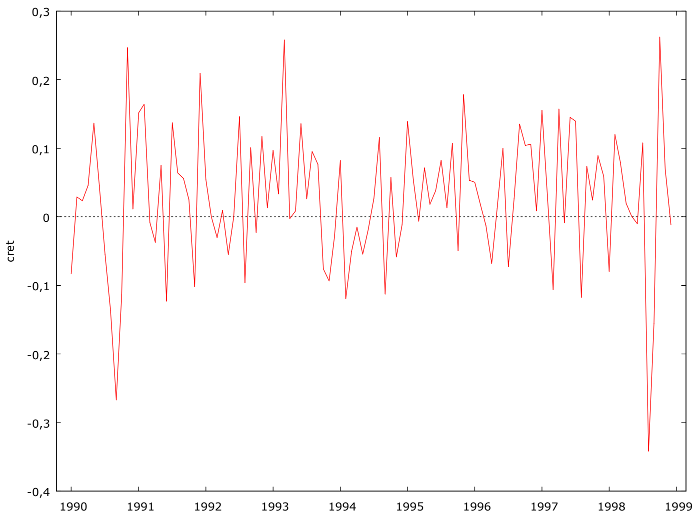



Figure 4.1. Monthly returns on Citigroup stock.

[1] Interestingly, the logarithmic return $R'_t = (\log(P_t) - \log(P_{t-1})) * 100$ gives practically the same values.

The returns oscillate rather regularly around some constant (which is greater than zero – this means that the returns are generally positive and therefore `cret` is increasing). It is a very simple time series, its future forecast is probably just this constant. One of the main objectives of this chapter is to learn how to forecast time series.

**4.2 example. Air passenger bookings**.

The number of international passenger bookings $Y_t$ or $AP_t$ (in thousands) per month on an airline (Pan Am) in the United States were obtained from the Federal Aviation Administration for the period 1949:1–1960:12 (this classic Box & Jenkins airline data is available as AP1.gdt or AP.txt in the data folder accompanying this course). The company used the data to predict future demand before ordering new aircraft and training aircrew.

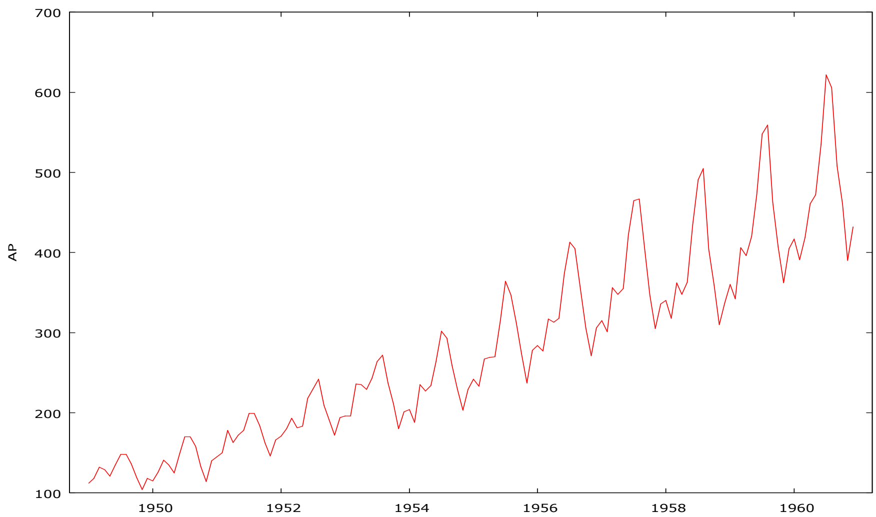



Figure 4.2. International air passenger bookings in the United States for the period 1949-1960

There are a number of features in the time plot of the air passenger data that are common to many time series. For example, it is apparent that the number of passengers travelling on the airline is increasing with time. In general, a systematic and deterministic change in a time series that does not appear to be periodic is known as a *trend* and denoted through $Tr_t$. The simplest model for a trend is a linear or exponential increase or decrease, and this is often an adequate approximation.

A repeating pattern within each year is known as *seasonal* variation (denonoted as $S_t$), although the term is applied more generally to repeating patterns within any fixed period, such as restaurant bookings on different days of the week. There is clear seasonal variation in the air passenger time series. At the time, bookings were highest during the summer months of June, July, and August and lowest during the autumn month of November and winter month of February.

It is clear that $Y_t$ is close to but not exactly equal to $Tr_t + S_t$. Any economic data is subject to some random disturbances or shocks $\varepsilon_t$, thus $Y_t = Tr_t + S_t + \varepsilon_t$ where $\varepsilon_t$ is some stationary (see below) series. Our purpose is to filter out these shocks and accurately estimate $Tr_t$ and/or $S_t$. Having done this, we shall be able to forecast $Y_t$.

**4.3 example. Quarterly exchange rate: GBP to NZ dollar.**

The trends and seasonal patterns in the previous two examples were clear from the plots. In addition, reasonable explanations could be put forward for the possible causes of these features. With financial data, exchange rates for example, such marked patterns are less likely to be seen, and different methods of analysis are usually required. A financial series may sometimes show a dramatic change that has a clear cause, such as a war or natural disaster. Day-to-day changes are more difficult to explain because the underlying causes are complex and impossible to isolate, and it will often be unrealistic to assume any deterministic component in the time series model.

The quarterly exchange rates for British pounds sterling to New Zealand dollars for the period 1991:1 to 2000:3 are shown in Fig. 4.3. The data (available as pounds_nz.dat) are mean values taken over quarterly periods of three months.

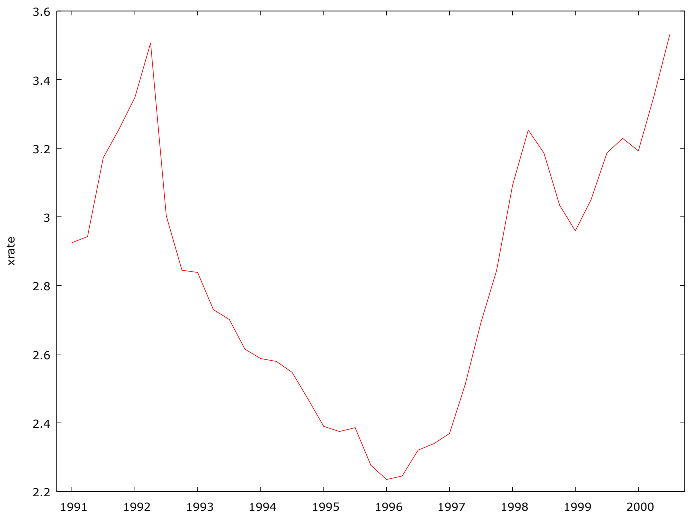



Figure 4.3. Quarterly exchange rates xrate for the period 1991–2000 (red)

The trend seems to change direction at unpredictable times rather than displaying the relatively consistent pattern of the air passenger series. Such trends have been termed stochastic trends to emphasise this randomness and to distinguish them from more deterministic trends like those seen in the previous examples. A mathematical model known as a random walk can sometimes provide a good fit to data like these and is discussed in the sequel.

## 4.2. Stationary Series

All time series can be divided into two big classes – (covariance or weak) stationary and nonstationary. Here is a short definition – a time series (or random process) is called **stationary** if it randomly but rather regularly (with more or less constant spread) fluctuates around its constant mean. Three examples of such series can be seen in Fig. 4.4.

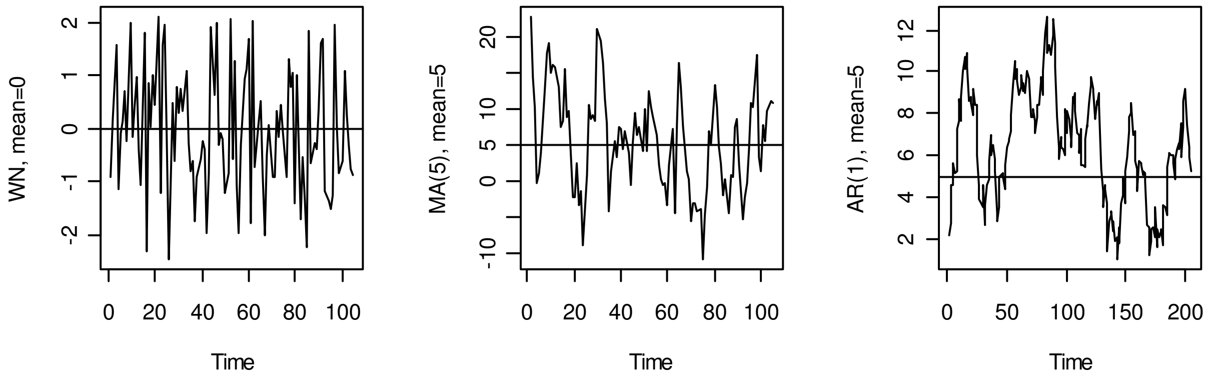



Figure 4.4. Three examples of stationary series; note that the third process (right) reverts to its mean more slowly that the previous two.

In Fig. 4.5 you can see four examples of nonstationary time series.

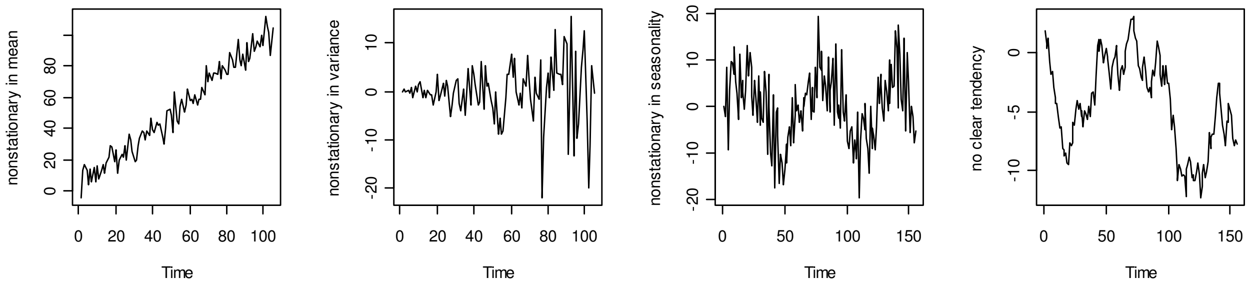



Figure 4.5. All four time series in this figure are not „rather regularly fluctuating around its constant mean"; these time series are not stationary

The simplest stationary random process, which is at the same time the main building block of all other stationary series, is the so-called **white noise** – this is a sequence of uncorrelated random variables with constant mean and constant variance (its graph is plotted in Fig. 4.4, left; note that the graph of the stock return, see Fig. 4.1, is quite similar to it). However, how can we know that the other two graphs of Fig.4.4 are not those of the WN? Two functions, ACF (autocorrelation function) and PACF (partial autocorrelation function), come to our rescue: if all the bars (except the zeroth in ACF) ar e within the blue band, the stationary process is WN (see Fig. 4.6).

The above graphical inspection is usually coupled with the Ljung-Box test designed to test the hypothesis $H_0$ : `cret` *is a WN*. To find the test's Q-statistics, open gretl, go to File * Open data * Sample file... * Ramanathan * data9-13, click on cret and right-click on correlogram:

```
Autocorrelation function for cret

LAG       ACF          PACF         Q-stat.  [p-value]

  1   -0.0346      -0.0346        0.1328   [0.716]
  2   -0.0773      -0.0786        0.8020   [0.670]
 ..............................................
 19   -0.0385      -0.0380        9.5634   [0.963]
```

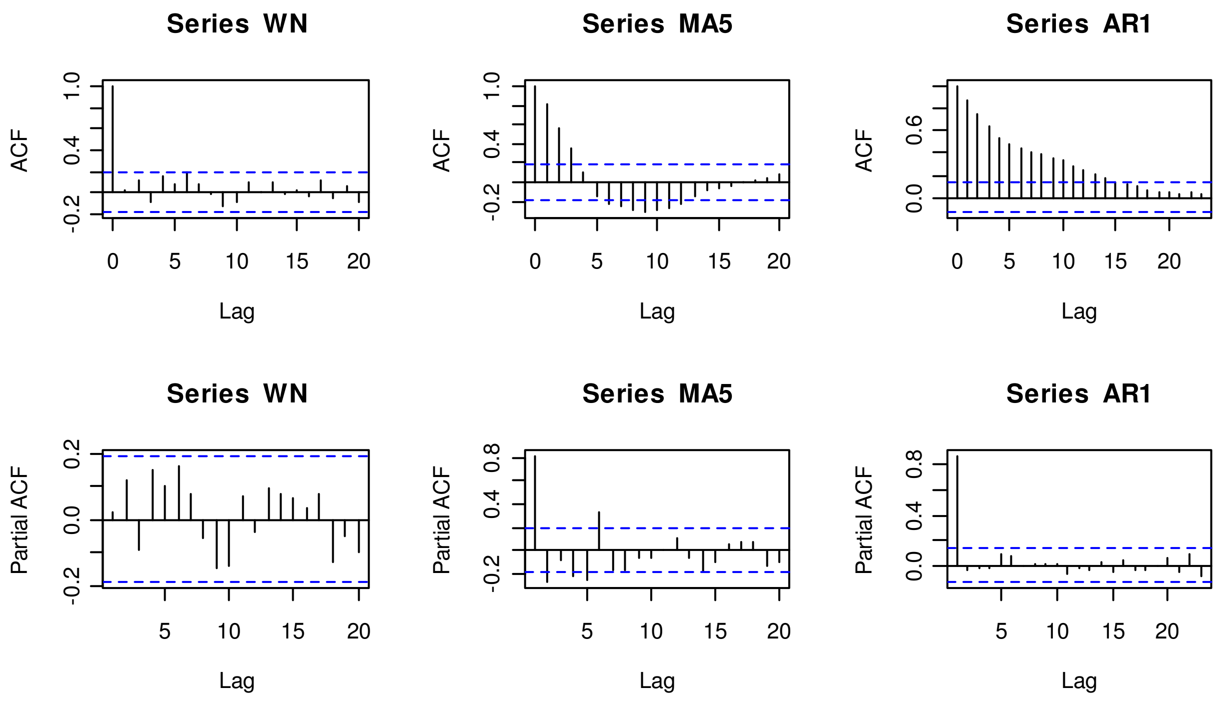



Figure 4.6. The time series WN is a white noise while the other two are not.

As always, the Q-stat. itself is not very important; instead, have a look at [p-value] of the $H_0$ - since all these numbers are greater than 0.05, there is no ground to reject the WN hypothesis. This is also confirmed by the correlogram (see Fig. 4.7):

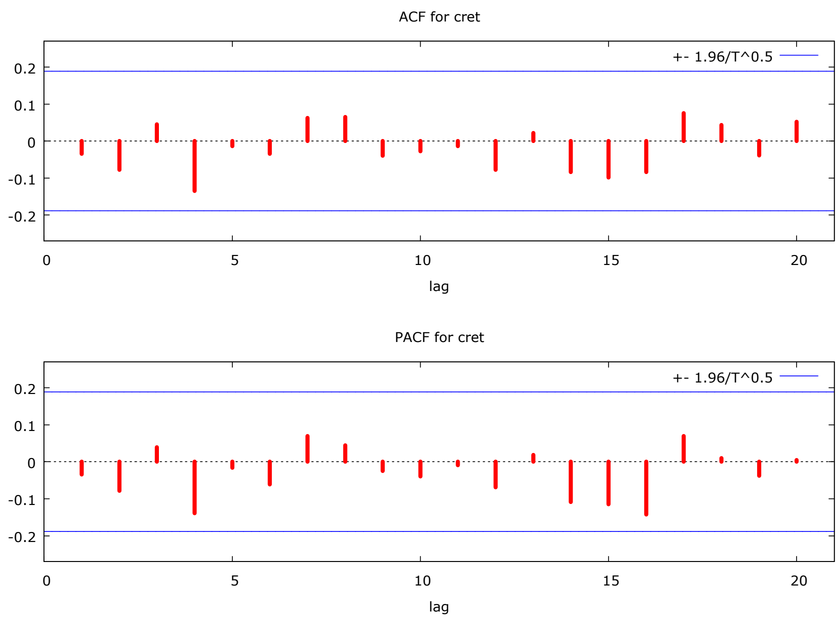



Figure 4.7. All the bars are inside the blue strips, thus, `cret` is WN

To decide whether the time series is stationary, examine its graph.
To decide whether a stationary time series is a white noise, examine its ACF and PACF.

The *forecast* of WN is trivial – as the past values do not correlate with the future, forecast does not depend on the series‘ past and equals just the (estimated) mean of `cret`. In gretl, go to Model * Time Series * ARIMA, choose `cret` as Dependent variable, insert 0‘s (zeros) in AR and MA boxes and click OK (you will see that the const (or the mean return) equals 0.0268522), in Model window choose Analysis * Forecasts..., insert 12 in the Number of observations to add box, click OK and you will get Fig. 4.8. As you can see, knowing past returns does not help to forecast them (the forecast is a constant).

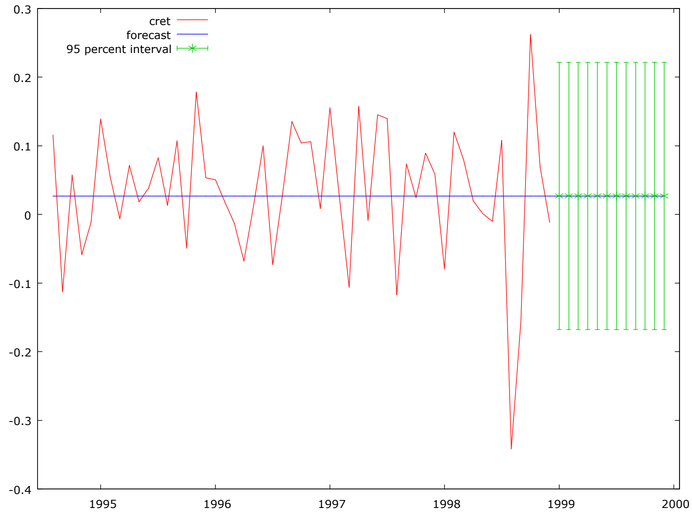



Figure 4.8. `cret` forecast together with its 95% confidence interval

Usually stationary processes are more complicated than WN, they incorporate possible dependence of the past and present. We shall analyse three types of stationary processes.

- **AR[2] processes**:

  the process $Y_t$ is called AR(1) process if it is described by the equation $Y_t = \alpha + \varphi_1 Y_{t-1} + \varepsilon_t, \varepsilon_t \sim WN$;
  the process $Y_t$ is called AR(2) process if it is described by the equation $Y_t = \alpha + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \varepsilon_t, \varepsilon_t \sim WN$ etc.

The <u>theoretical</u> (or population) ACF and PACF of AR(1) are depicted in Fig. 4.9 (note one non-zero bar in PACF). The following rule helps to classify the time series as AR:

| |
|---|
| If in <u>sample</u> PACF of a time series $p$ bars are outside the blue band and ACF gradually declines, the time series is probably AR(p) |

[2] AR=AutoRegressive

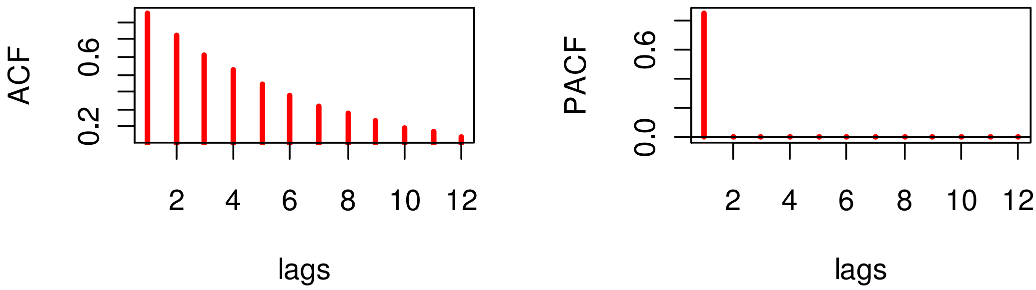



Figure 4.9. Theoretical AC function (left) and PAC function (right) for the AR(1) process with $\varphi_1 = 0.85$.

AR process is stationary and has the mean reverting[3] property if its coefficients satisfy certain conditions. For example, the AR(1) process $Y_t = \alpha + \varphi_1 Y_{t-1} + \varepsilon_t$ is stationary if $|\varphi_1| < 1$. Note that if $\varphi_1$ is close to zero, the process is almost the WN (see Fig. 4.10, left) but if it is close to 1, the trajectories are more persistent (have the inertia property) and revert to zero not so fast (see Fig. 4.10, right).

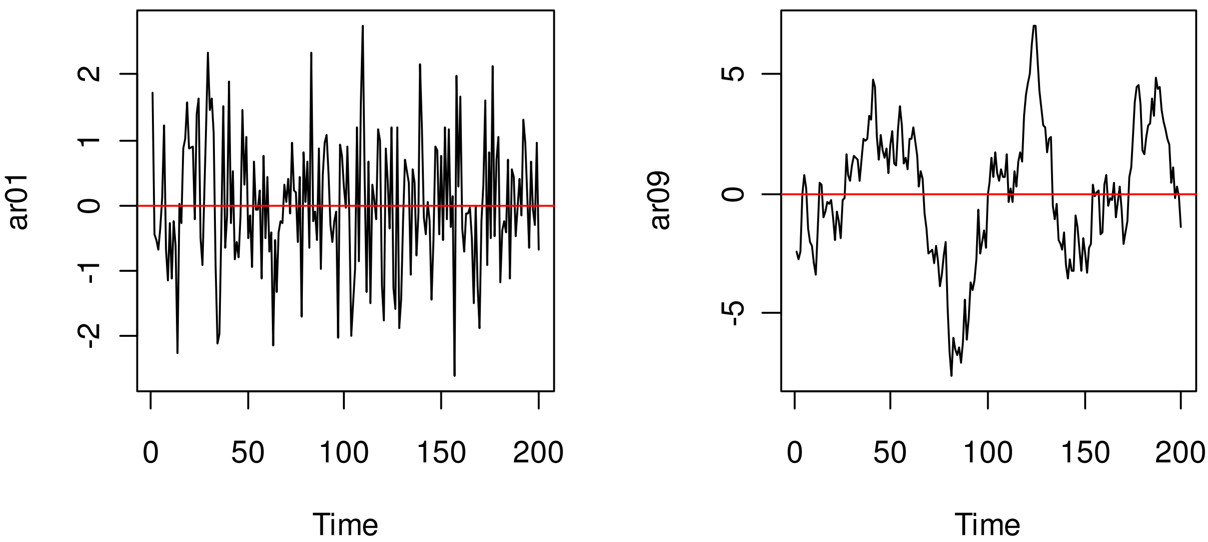



Figure 4.10. One trajectory of the AR(1) process ($\varphi_1 = 0.1$, left and $\varphi_1 = 0.9$, right)

**4.4 example.** Open gretl and import quarterly data of US unemployment from 1948:01 through 1978:01 in unemp.txt. Plot the time series (it seems to be stationary) and examine the correlogram (it indicate that the series is probably AR(2), see Fig. 4.11).

---

[3] The property means that if the AR(1) process $Y_t = \alpha + \varphi_1 Y_{t-1} + \varepsilon_t$ has no more shocks, i.e., $\varepsilon_t \equiv 0,\ t \geq T$, the path (of its forecast) tends to the mean.

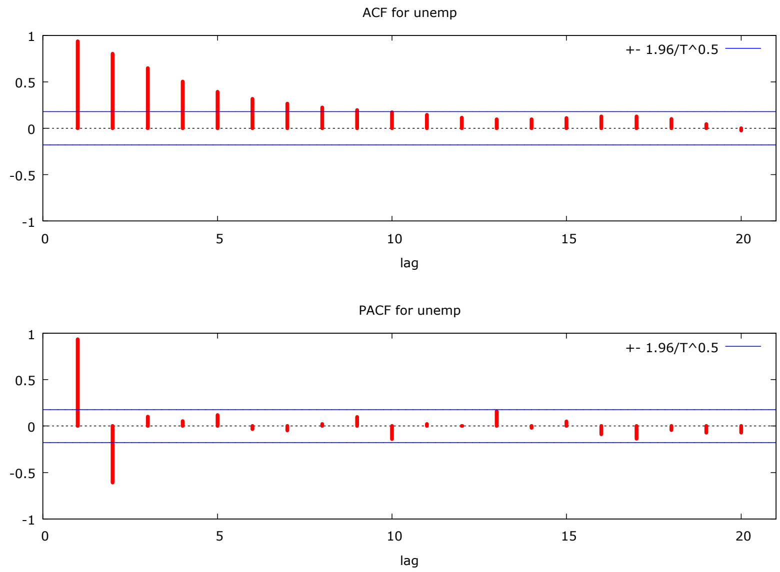



Figure 4.11. Two bars in PACF peeps out from the blue band, while ACF gradually declines.

To forecast unemp for two years (8 quarters), go to Model * Time Series * ARIMA, choose unemp as Dependent variable, insert 2 in AR and 0 in MA boxes and click OK, in Model window choose Analysis * Forecasts..., insert 8 in the Number of observations to add box, click OK and you will see the forecast in Fig. 4.12. As it ought to be for stationary series, the forecast approaches the mean value of `unemp`, but our procedure allows us to quantify the rate of approach.

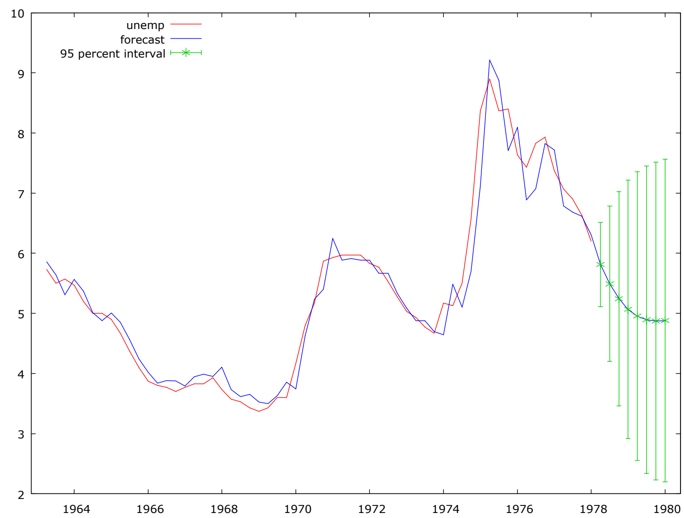



Figure 4.12. Some later historical data of `unemp` together with its 8-quarters-forecast (the forecast tends to the mean)

- **MA[4] processes**:

  the process $Y_t$ is called MA(1) process if it is described by the equation $Y_t = \alpha + \varepsilon_t + \theta_1 \varepsilon_{t-1}, \varepsilon_t \sim WN$;
  the process $Y_t$ is called MA(2) process if it is described by the equation $Y_t = \alpha + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2}, \varepsilon_t \sim WN$ etc.

The theoretical (or population) ACF and PACF of MA(3) are depicted in Fig. 4.13 (note three non-zero bars in ACF). The following rule helps to classify the time series as MA:

If in sample ACF of a time series $q$ bars are outside the blue band and PACF gradually declines, the time series is probably MA(q)

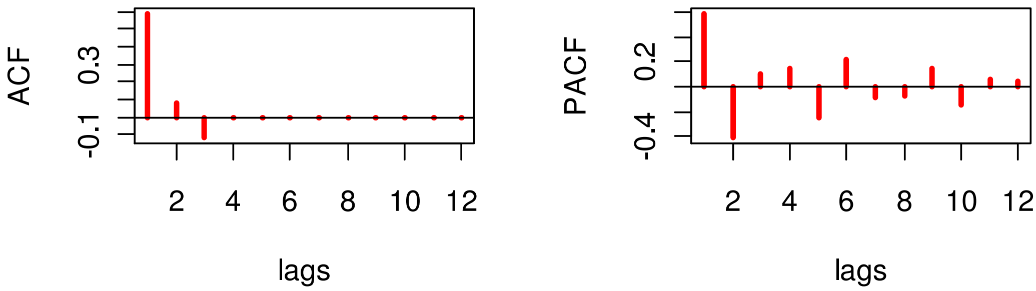



Figure 4.13. The theoretical correlogram of the MA(3) process with $\theta_1 = 1.2$, $\theta_2 = 0.65$, $\theta_3 = -0.35$ (note that its ACF cuts off at $t = 3$ and PACF decays gradually).

- **ARMA[5] processes**:

  the process $Y_t$ is called ARMA(p,q) process if it is described by the equation $Y_t = \alpha + \varphi_1 Y_{t-1} + ... + \varphi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + ... + \theta_q \varepsilon_{t-q}, \varepsilon_t \sim WN$.

If both sample ACF and PACF gradually decline, the time series is probably ARMA(p,q) with some $p$ and $q$ which are still to be estimated.

**4.5 example.** Figure 4.14 shows the plot of of artificial annual time series dated from 1 to 200 (the data is available as .../data/arma11.txt).

[4] MA=MovingAverage
[5] ARMA=AutoRegressiveMovingAverage

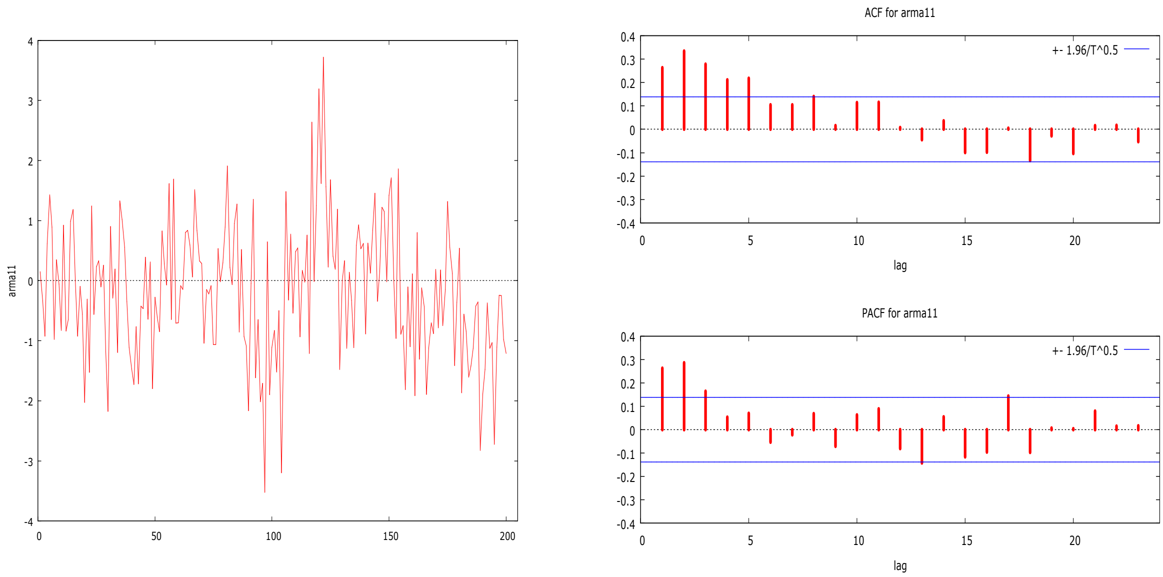
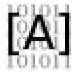



Figure 4.14. The graph of `arma11` (left) and its correlogram (right)

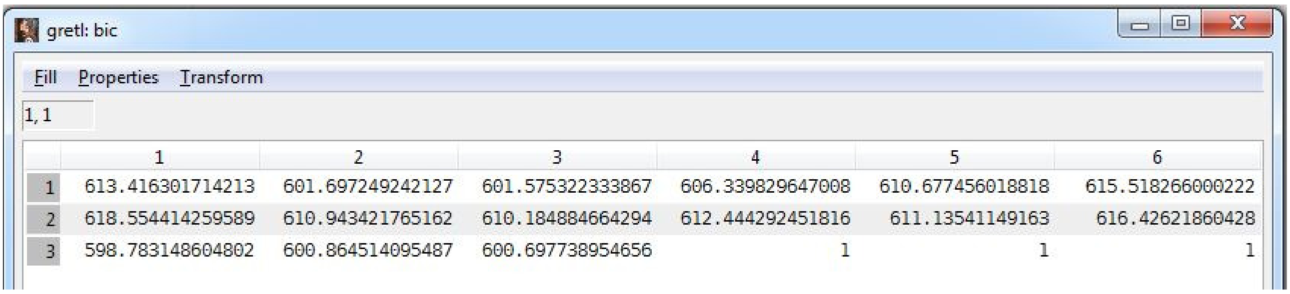
The correlogram is rather complicated: `arma11` may well be AR(3), MA(5), or ARMA with not quite evident $p$ and $q$. We shall test many models and choose the „best“, namely, the one with the smallest value of the Schwarz criterion. Copy and paste the following script to gretl‘s New script window (open the window by clicking on the second from the left icon, see bottom line in Fig. 1.1):

```
matrix bic = ones(3,6)          # Create an auxiliary matrix (called bic)of 1's

arima 1 0 0 ; arma11            # Model arma11 as ARIMA(1,0,0)=AR(1)
genr bic[1,1] = $bic            # Fill in the (1,1) element of the matrix
                                # with the model‘s Schwarz coefficient
arima 2 0 0 ; arma11            # Continue...
genr bic[1,2] = $bic
arima 3 0 0 ; arma11
genr bic[1,3] = $bic
arima 4 0 0 ; arma11
genr bic[1,4] = $bic
arima 5 0 0 ; arma11
genr bic[1,5] = $bic
arima 6 0 0 ; arma11
genr bic[1,6] = $bic
arima 0 0 1 ; arma11            # Model arma11 as ARIMA(0,0,1)=MA(1)
genr bic[2,1] = $bic            # Fill in the second line of the matrix
arima 0 0 2 ; arma11            # Continue...
genr bic[2,2] = $bic
arima 0 0 3 ; arma11
genr bic[2,3] = $bic
arima 0 0 4 ; arma11
genr bic[2,4] = $bic
arima 0 0 5 ; arma11
genr bic[2,5] = $bic
arima 0 0 6 ; arma11
genr bic[2,6] = $bic
arima 1 0 1 ; arma11            # Model arma11 as ARMA(1,1)
genr bic[3,1] = $bic            # Continue...
arima 2 0 1 ; arma11
genr bic[3,2] = $bic
arima 1 0 2 ; arma11
genr bic[3,3] = $bic
```

To run the script, click on the pinion (i.e., the right-most) icon in


The script produces a matrix filled with respective bic values. We shall choose the model with the smallest value of Schwarz‘s criterion (to see the matrix, click on the fourth from the left icon in the bottom row, see Fig. 1.1, and click on the  icon:

bic




| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | 613.416301714213 | 601.697249242127 | 601.575322333867 | 606.339829647008 | 610.677456018818 | 615.518266000222 |
| 2 | 618.554414259589 | 610.943421765162 | 610.184884664294 | 612.444292451816 | 611.13541149163 | 616.42621860428 |
| 3 | 598.783148604802 | 600.864514095487 | 600.697738954656 | 1 | 1 | 1 |

The smallest value of Schwarz‘s criterion in this matrix corresponds to the (3,1)-element (it equals 598.78), i.e., we shall describe arma11 as ARMA(1,1) process. Note that the second best is ARMA(1,3) with 600.69.

**Exercise 4.1**. Assume that we choose the model MA(3) (it is the best among MA models, isn‘t it?). Create this model. Do the residuals of this model make a WN? (in the model window go to Graphs). What about residuals of the ARMA(1,1) model? Forecast arma11 for one year with ARMA(1,1).

To choose the right model for a stationary time series, estimate several models „close“ to that recommended by correlogram. Choose the one with the smallest value of Akaike or Schwarz criterion, provided its residuals make a WN

One of the possibilities to assess a model is to stop the process sometime before the final date, create a model and to compare its prediction with the real data. This is illustrated in the next exercise.

**Exercise 4.2**. Import the quarterly time series .../data/caemp.txt (this is seasonally adjusted Canadian index of employment, 1962:1-1995:4). The series displays no trend, and, of course, it displays no seasonality because it is seasonally adjusted. It does, however, appear highly serially correlated as it evolves in a slow, persistent fashion. 1) Can you prove that it is not a WN?

Now narrow the time range of the caemp to 1962:1 – 1993:4 (in order to do this, go to Sample * Set range... and set the above-mentioned range). 2) Draw correlogram. What is the right model? Why? Can you confirm that the best model is AR(2)? (use a program similar to that of 4.5 example). 3) Use the AR(2) model to predict caemp until the original end of 1995:4 and compare the forecast with the real data. 4) Use the MA(6) model to the same purpose and find which of the two models has better forecasting properties? ◄◄

The forecast of any stationary time series tends to its mean (but the exact manner of convergence depends on the type of the process and its coefficients).

## 4.3. TS Series

We have already mentioned that some time series may be expressed as $Y_t = Tr_t + S_t + \varepsilon_t$ where $Tr_t$ is a deterministic trend, $S_t$ deterministic seasonal part, and $\varepsilon_t$ is a random stationary series (such a series $Y_t$ is called a trend stationary (TS) series). Some 30 years ago most economists though that all the time series are TS series. However, later on it appeared that most economic or financial series are more complicated (see next section).

**4.6 example.** We shall analyse the AP.txt time series (see 4.2 example). Our purpose is to *decompose* the series, that is, capture trend and seasonal part, classify the residual process and then forecast each part of the series one year ahead. For the easy of exposition, we present the relevant script which, after importing the AP.txt data, can be copied to gretl's script window and executed with the `Run` button:

```
# The trend seems to be parabolic, therefore we'll include squares of time
# Generate dates (as 1,2,3,...) and their squares
genr time
genr t2=time*time
# Generate monthly dummies (to capture seasonal effects)
genr dummy
# Extract parabolic (or quadratic or square) trend only
ols AP 0 time t2
# Fit historical values
fcast yhata1
# We shall forecast bookings for the next 12 months
# Extend range
addobs 12
# Fit historical data and add forecast for the next 12 months
smpl 1949:1 1961:12
fcast 1949:1 1961:12 yhata2
# Estimate trend and seasonal effects (use historical data only)
smpl 1949:1 1960:12
ols AP 0 time t2 dm2 dm3 dm4 dm5 dm6 \
dm7 dm8 dm9 dm10 dm11 dm12
fcast yhatb1
# Fit historical data and add forecast for the next 12 months
addobs 12
smpl 1949:1 1961:12
fcast 1949:1 1961:12 yhatb2
```

Select `AP`, `yhata2`, and `yhatb2` and plot them all – the amplitude of the forecast `yhatb2` is always the same while that of AP is increasing (see Fig. 4.15, left) which means that our model is unsatisfactory.

If the fluctations of the time series $Y_t$ are increasing together with its level, create a model for $\log Y_t$

```
# Create the model for log(AP)
smpl 1949:1 1960:12
logs AP
# Extract trend and seasonal effect
ols l_AP 0 time t2 dm2 dm3 dm4 dm5 dm6 \
dm7 dm8 dm9 dm10 dm11 dm12
fcast yhatc1
# Save the residuals
series uhatc1 = $uhat
# Forecast next year bookings in logs
addobs 12
```

```
smpl 1949:1 1961:12
fcast 1949:1 1961:12 yhatc2
```

Now select `l_AP` and `yhatc2` and plot them both (Fig. 4.15, right) – the model seems to be quite satisfactory[6].

The last step is to get back to $AP_t$ by using the formula $\widehat{AP_t} = \exp(\widehat{\log(AP_t)} + \hat{\sigma}_u^2 / 2)$ [7]:

```
series yhatd = exp(yhatc2+$sigma/2)
```

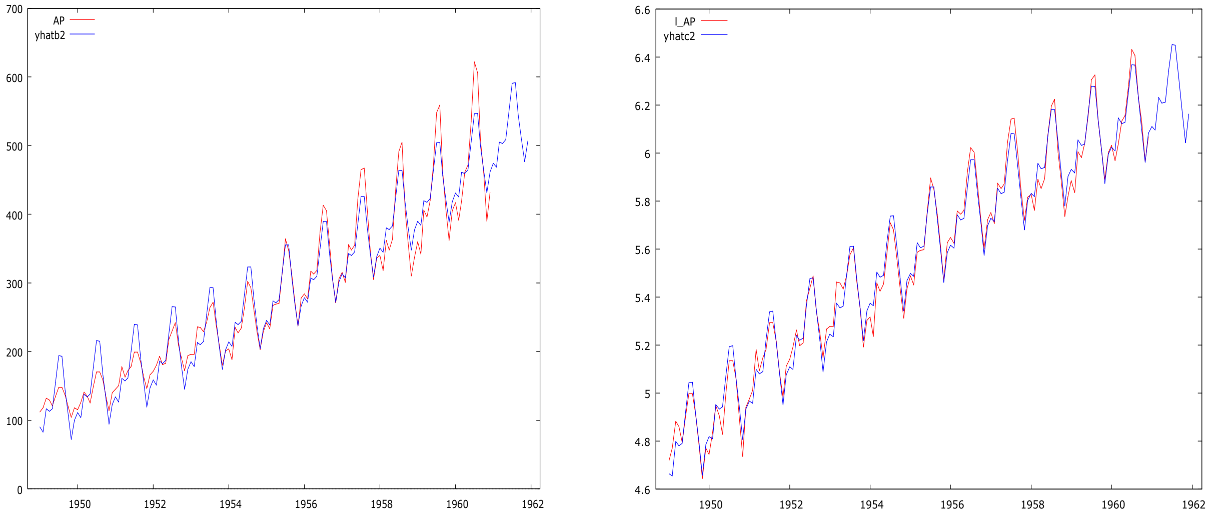



Figure 4.15. Trend and seasonal component for `AP` (left) and `l_AP` (right)

To get the feeling of the accuracy of the forecast, one may now select `AP` and `yhatd` and to plot both series. ◀◀

**Exercise 4.3**. The data available as shampoo.txt is a monthly sales of shampoo over the period 2000:1 – 2002:12 (the data has no seasonal component). From the menu bars, do the following: 1) Extract the square time trend. 2) Forecast the sales of shampoo one year (not one month!) ahead. ◀

In order to decompose a TS series (i.e., to extract $Tr_t$, $S_t$, and $\varepsilon_t$), one can also use another smoothing or filtering procedures collected in GRETL in the Variablel Filter section. For example, select the `AP` variable of the 4.6 example, go to Variablel Filterl Simple moving average, choose Number of observations in average **3**, and check Centered. For every *t*, this procedure estimates the *moving average* (here, with equal weights $w_{-1} = w_0 = w_1 = 1/3$): $w_{-1}AP_{t-1} + w_0 AP_t + w_1 AP_{t+1} = (1/3)AP_{t-1} + (1/3)AP_t + (1/3)AP_{t+1}$ (see Figure 4.16, left). In order to remove the seasonal part of `AP`, replace the number **3** by **12** (i.e., the period; since **12** is an even number, the weights $w_i$ now are more complicated). Another possibility is to use the exponential moving average etc. The

[6] In fact, we should now test its residuals for WN and, if neccesary, to correct the model for serial correlation (in GRETL this is called an ARMAX model); we shall skip this analysis.
[7] To find the explanation of `$sigma`, go to gretl's Help * Function reference (the term `$sigma/2` is necessary to correct the bias of the fit).

common drawback of the most of filtering methods is that they do not allow to forecast the time series (this remark does not apply to exponential smoothing which can be extended into the future).

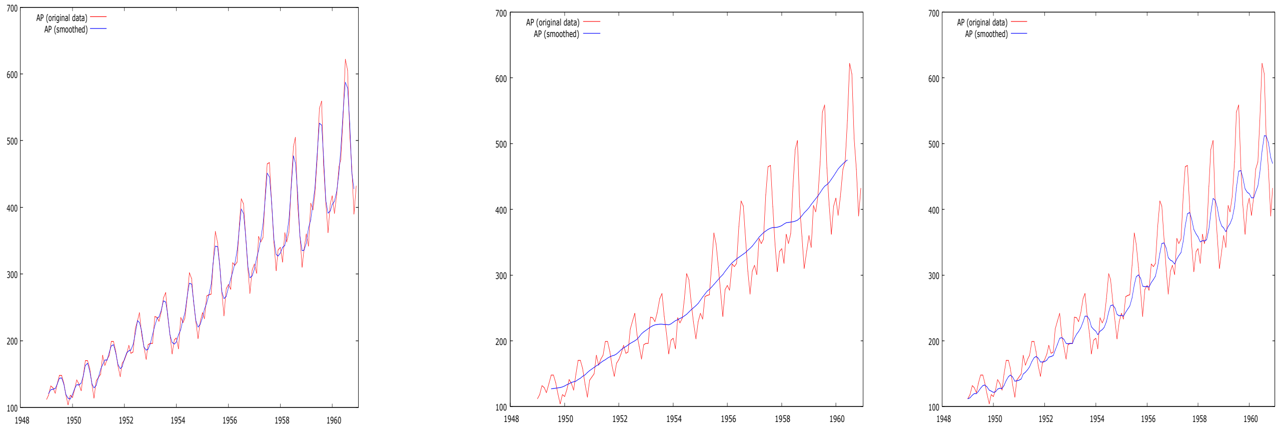



Figure 4.16. 3-term and 12-term centered moving average of AP (left and center) and exponential moving average (weight on current observation 0.2, right).

## 4.4. DS Series

It has already been mentioned that the AR(1) process $Y_t = \mu_0 + \varphi_1 Y_{t-1} + \varepsilon_t$ is stationary (and therefore has the mean reverting property) <u>if</u> $|\varphi_1| < 1$. Note that the properties of the process (for example, its forecast) change considerably if the above stationarity condition is violated. In the concrete, if $\varphi_1 = +1$, we say that $Y_t$ has a *unit root*. More specifically, the unit root process $Y_t = Y_{t-1} + \varepsilon_t$ is called a *random walk* (RW); $Y_t = \mu_0 + Y_{t-1} + \varepsilon_t$, $\mu_0 \neq 0$, is called a RW with a *drift* $\mu_0$. Note that the differences of random walk, i.e., $\Delta Y_t = Y_t - Y_{t-1} = (\mu_0 +)\, \varepsilon_t$ form a stationary process. In general,

A series $Y_t$ is called Difference Stationary (DS) if it is 1) not a TS series but 2) its differences $\Delta Y_t$ make a stationary process.

**4.7 example.** If one tries to extract a trend from a (nonstationary!) random walk $Y_t = \varepsilon_1 + \ldots + \varepsilon_t$, $\varepsilon_t \sim WN$, using, for example, a moving average procedure: $\tilde{Y}_t = (Y_{t-1} + Y_t + Y_{t+1})/3 = ((\varepsilon_1 + \ldots + \varepsilon_{t-1}) + (\varepsilon_1 + \ldots + \varepsilon_{t-1} + \varepsilon_t) + (\varepsilon_1 + \ldots + \varepsilon_{t-1} + \varepsilon_t + \varepsilon_{t+1}))/3 = \varepsilon_1 + \ldots + \varepsilon_{t-1} + (\varepsilon_t + \varepsilon_{t+1})/3$, this new series $\tilde{Y}_t$ will be more smooth that the original series, but it will be random (every walk will have its <u>own</u> *random* "trend", see Fig. 4.17). Note that differences $\Delta Y_t = \varepsilon_t$ will be stationary. ◄◄

The process $Y_t$ whose first differences make a stationary ARMA(p,q) process is called an ARIMA(p,1,q) process. If the process itself is a stationary ARMA(p,q) process (no differencing is needed to make it stationary), it can also be called ARIMA(p,0,q).

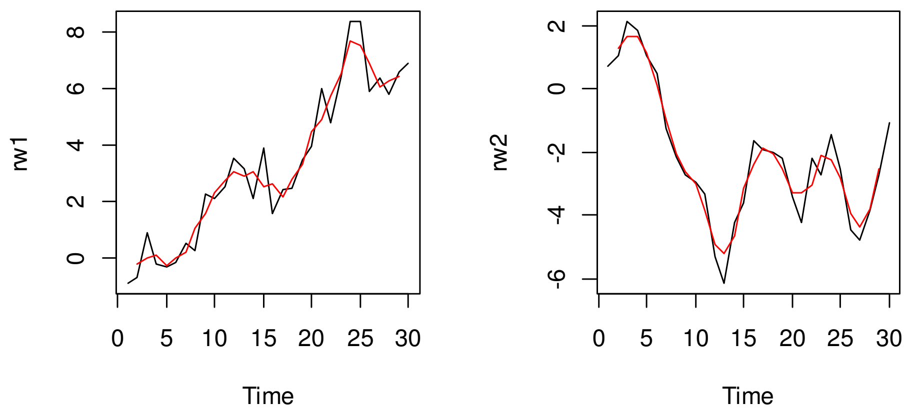



Figure 4.17. Two paths of a random walk and their smoothed versions (the red „trends" are random).

Often, it is not so easy to distinguish whether a given series is described by a random walk $Y_t = \mu_0 + 1 \cdot Y_{t-1} + \varepsilon_t$ or by a stationary AR(1) process $Y_t = \mu_0 + \varphi_1 \cdot Y_{t-1} + \varepsilon_t$ with $|\varphi_1| < 1$. However, in order to test the crucial hypothesis $H_0 : \varphi_1 = 1$ versus $H_1 : \varphi_1 < 1$, we cannot use ols with its usual $t-$ratio $\hat{\varphi}_1 / \widehat{se(\hat{\varphi}_1)}$ and respective Student's $p-$value from the GRETL regression table (because <u>if</u> $H_0$ <u>is true</u>, $t-$ratio has <u>another</u>, Dickey-Fuller's, distribution).

Rewrite $Y_t = \mu_0 + \varphi_1 Y_{t-1} + \varepsilon_t$ as $\Delta Y_t = \mu_0 + \rho Y_{t-1} + \varepsilon_t$, where $\rho = \varphi_1 - 1$, and add a possible trend to generalize it to $\Delta Y_t = \mu_0 + \mu_1 t + \rho Y_{t-1} + \varepsilon_t$ (Fig. 4.18 shows that this equation allows to describe quite different behavior of random series).

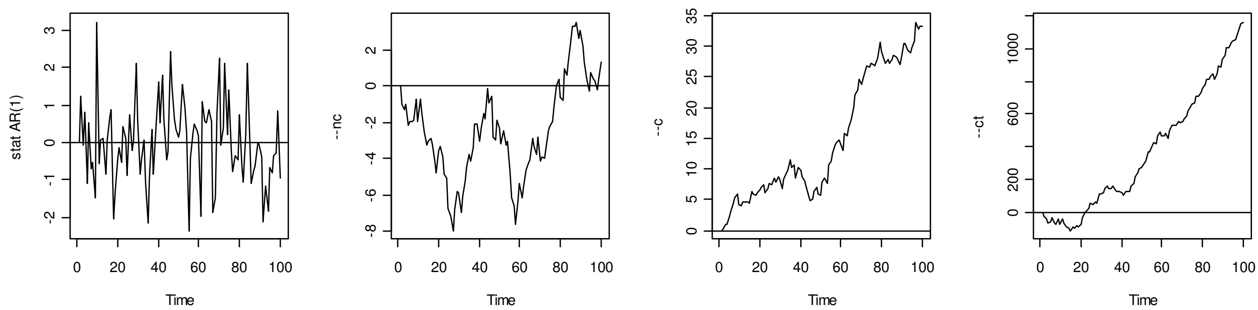



Figure 4.18. Modeled stationary AR(1) process ($|\varphi_1| < 1, \mu_0, \mu_1 \neq 0$, left), non-stationary random walk ($\varphi_1 = 1, \mu_0 = \mu_1 = 0$, second from the left), non-stationary random walk with positive (or upward) drift ($\varphi_1 = 1, \mu_0 > 0, \mu_1 = 0$, second from the right), and non-stationary random walk with ever increasing drift ($\varphi_1 = 1, \mu_0$ and $\mu_1 \neq 0$, right )

The hypothesis $H_0 : Y_t$ *has a unit root* is equivalent to $H_0 : \varphi_1 = 1$ or, what is the same, $H_0 : \rho = 0$

Thus, we want to test $H_0: \rho = 0$ against $H_1: \rho < 0$ (that is, unit root against stationarity). In order to do this, we shall still generalize (by adding more lags) the previous equation to

$$\Delta Y_t = \mu_0 + \mu_1 t + \rho Y_{t-1} + \sum_{i=1}^{p} \gamma_i \Delta Y_{t-i} + \varepsilon_t \qquad (4.1)$$

and use the ADF (*augmented Dickey-Fuller*) test. We shall explain the procedure by means of

**4.8 example.** The file lGNP.txt contains annual data of logged GNP (USA, 1950-1993).

**1)** It seems that `lGNG` could be well described (see red curve in Fig. 4.19, left) as a **TS process** $\texttt{lGNP}_t = \alpha + \beta t + \varepsilon_t$ (where $\varepsilon_t$ is a stationary process) therefore we shall extract the linear trend first. It can be done with the ols procedure but now it is more convenient to use the ARIMA model. We shall need the time trend, therefore create it first: go to Add| Time trend. Now go to Model| Time series| ARIMA..., choose `lGNP` as dependent, `time` as independent and insert 0 as AR and MA order. The model obtained, i.e., $\texttt{lGNP}_t = 14.22 + 0.0321\ t$ [8], can be analyzed from different perspectives (for example, go to Graphs in the model window), but what is important to us is to test whether `lGNP` is a TS series, i.e., whether the <u>residuals</u> of the model are stationary or constitute RW. This is where we shall use the ADF test: in the model window, click on Save| Residuals and denote them as `uhat1`, go to the gretl window, select `uhat1`, choose Variable| Unit root tests| Augmented Dickey-Fuller test and mark „test without constant"[9] box – the procedure will automatically choose $p$ in (4.1) and also estimate the $p$–value of the hypothesis $H_0$: `uhat1` *has a unit root* test (it equals 0.08 (>0.05), therefore we have no ground to reject $H_0$). Thus, `lGNP` is not TS (residuals are not stationary) and it is illegal to use the model $\texttt{lGNP}_t$ = $14.22 + 0.003\,t$ for forecasting. Nevertheless, to get a feeling of the forecast, in the model window go to Analysis| Forecasts... and add 7 observations – what you get is shown in Fig. 4.20, left.

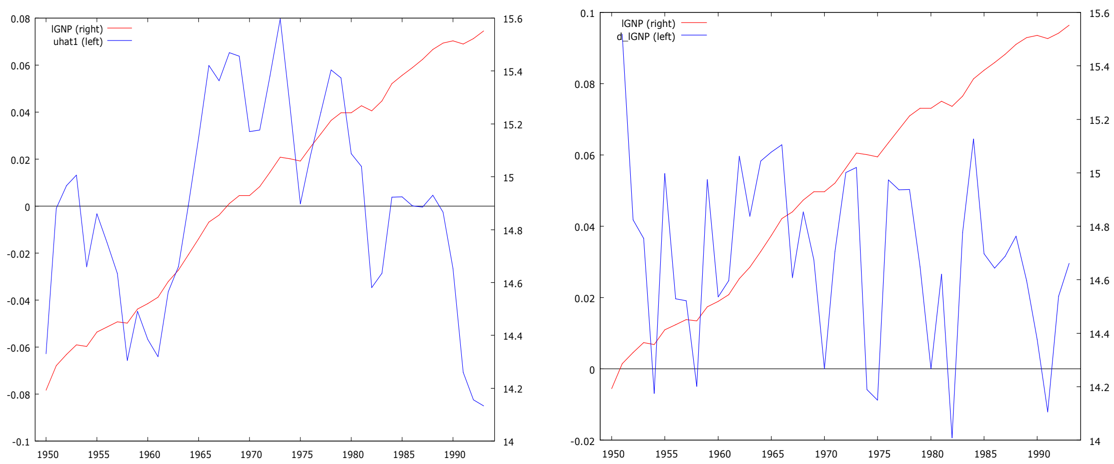



Figure 4.19. The graph of non-stationary `lGNP` and its RW residuals `uhat1` (left) and the graph of non-stationary `lGNP` and its stationary differences (right).

[8] It means that the average growth of the GNP was 3.2% during these years.
[9] We choose the box to check according to the alternative; in our case, it is „`uhat1` is a stationary AR process around a zero constant" (we choose such an alternative because `uhat1` has a zero mean by construction).

**2)** Now we shall test whether `lGNP` can be presented as **DS process** – select `lGNP`, go to Variable| Unit root tests| Augmented Dickey-Fuller test| check the „with constant and trend"[10] box and press OK. The $p-$value equals 0.78 (>0.05), thus we do not reject the unit root (or RW with a drift or $H_0$) hypothesis, i.e., `lGNP` must be described as $\texttt{lGNP}_t - \texttt{lGNP}_{t-1} = \mu_0 + \varepsilon_t$. To find the coefficient $\mu_0$, in GRETL's window go to Model| Time series| ARIMA..., choose `lGNP` as dependent variable, AR and MA orders set equal to 0, and Difference to 1 – the estimate of $\mu_0$ equals 0.0316. Both coeficients (growth rates) are only marginally different, more important are the differences in forecasts (in DS case, the errors are ever increasing, see Fig. 4.20, right; also, compare the 2000-forecasts – 15.8616 in TS case and 15.7730 in DS case). The bottom line – use the DS model.

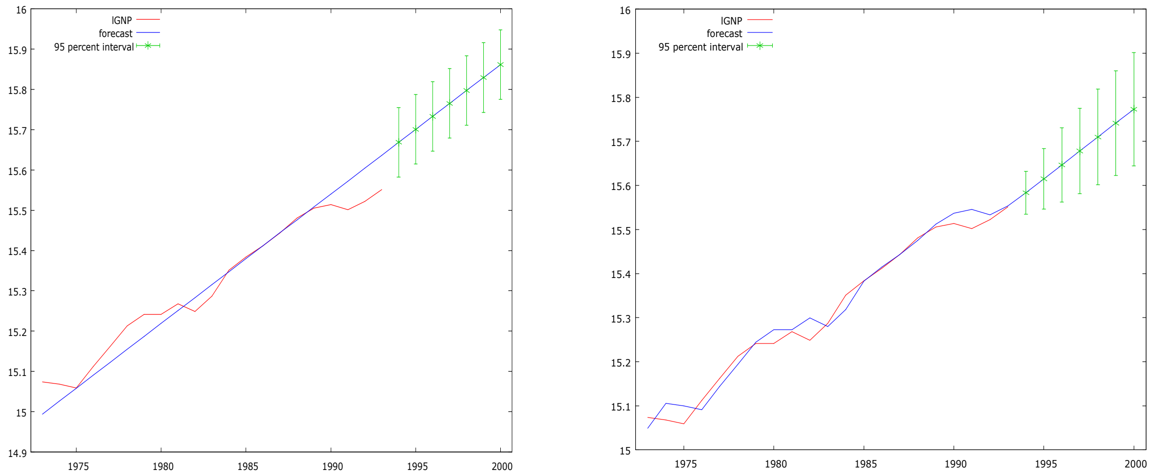



Figure 4.20. The 7-years-ahead forecast with the TS model (left); the same forecast with DS model (right)

> If the process under consideration is a DS process, its forecast is a horizontal line (if there is no drift) and a straight line with the slope equal to the drift, otherwise.

**4.9 example.** The quarterly exchange rates `xrate` for British pounds sterling to New Zealand dollars for the period 1991:1 to 2000:3 are available in pounds_nz.dat.

1) Describe `xrate` as a TS series with first, second, and third order polynomial trends. Which model is the best (according to your graphs and AIC)?

```
Model 1: OLS, using observations 1991:1-2000:3 (T = 39)
Dependent variable: xrate

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const      2.72423       0.125348     21.73     1.16e-022 ***
  time       0.00495126    0.00546198    0.9065   0.3705

Log-likelihood       -16.97411   Akaike criterion     37.94822
rho                   0.963689   Durbin-Watson        0.129688
```

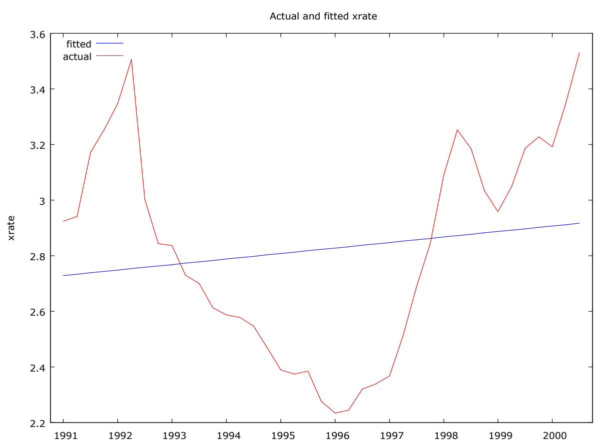
[10] We choose the box to check according to the alternative; in our case it is $H_1$ : `lGNP` *is stationary AR process around a trend*; recall that $H_0$ is „`lGNP` *is RW with a drift*".

```
Model 2: OLS, using observations 1991:1-2000:3 (T = 39)
Dependent variable: xrate

             coefficient   std. error    t-ratio    p-value
  ---------------------------------------------------------
  const        3.48144     0.107322      32.44     3.02e-028 ***
  time        -0.105861    0.0123736     -8.555    3.37e-010 ***
  sq_time      0.00277030  0.000300003    9.234    4.99e-011 ***

Log-likelihood        6.708823   Akaike criterion    -7.417646
Schwarz criterion    -2.426961   Hannan-Quinn        -5.627031
rho                   0.728715   Durbin-Watson        0.415566

Model 3: OLS, using observations 1991:1-2000:3 (T = 39)
Dependent variable: xrate

             coefficient    std. error    t-ratio    p-value
  ----------------------------------------------------------
  const        3.42550      0.151527      22.61     1.82e-022 ***
  time        -0.0900682    0.0323927     -2.781    0.0087    ***
  sq_time      0.00179565   0.00186905     0.9607   0.3433
  cu_time      1.62440e-05  3.07386e-05    0.5285   0.6005

Log-likelihood        6.863797   Akaike criterion    -5.727595
rho                   0.741839   Durbin-Watson        0.411909
```

The coefficient of $(time^3 =)$ `cu_time` in Model 3 is insignificant, Akaike criterion is minimum in Model 2, and the quadratic trend in Fig. 4.21 is much better than the linear one, therefore we choose Model 2:

```
^xrate = 3.48 - 0.106*time + 0.00277*sq_time
        (0.107) (0.0124)     (0.000300)
```

However, note that `rho` in Model 2 (in $\hat{\varepsilon}_t = (rho\,\hat{\varepsilon}_{t-1} =)\, 0.729\,\hat{\varepsilon}_{t-1}$) is far from 0 and Durbin-Watson‘s statistics of residuals (=0.416) is far from 2, thus `xrate` is most probably not a TS series and using parabola for prediction is incorrect.



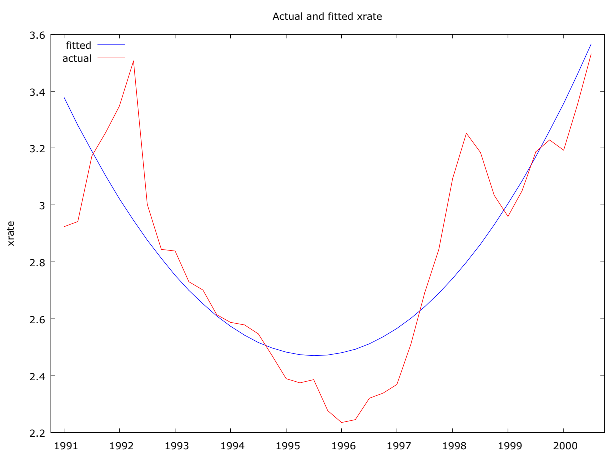




Figure 4.21. `xrate` and linear trend (left) and `xrate` and quadratic trend (right)

2) Nevertheless, to start with, we use Model 2 to predict `xrate` 8 quarters ahead: in GRETL main window, go to Data| Add observations| 8| OK; now go to Model 2 window| Analysis| Forecasts...| OK (see Fig. 4.22, left).

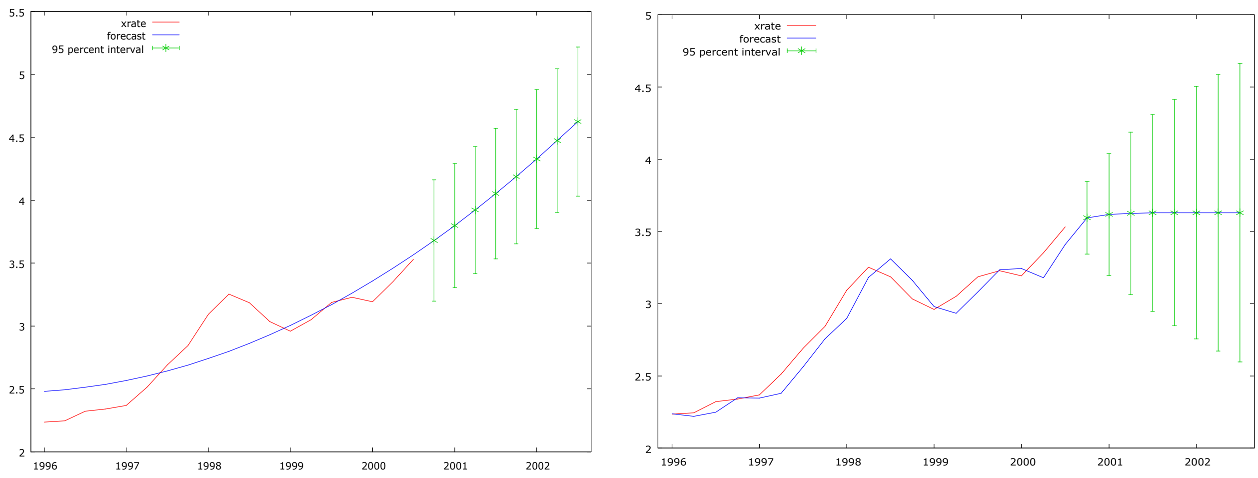



Figure 4.22. Two forecasts: quadratic model (left) and ARIMA(1,1,0) model (right)

3) Now we shall test whether `xrate` is a DS process, i.e., we shall test it for a unit root. In the main GRETL window, select `xrate` and go to Variable| Unit root tests| Augmented Dickey-Fuller test| check the „show regression results“ box| OK.

```
Augmented Dickey-Fuller test for xrate
including one lag of (1-L)xrate (max was 9)
sample size 37
unit-root null hypothesis: a = 1

*********************************

   test with constant
   model: (1-L)y = b0 + (a-1)*y(-1) + ... + e
   1st-order autocorrelation coeff. for e: 0.008
   estimated value of (a - 1): -0.0582389
   test statistic: tau_c(1) = -0.950578
   asymptotic p-value 0.7725

Augmented Dickey-Fuller regression
OLS, using observations 1991:3-2000:3 (T = 37)
Dependent variable: d_xrate

              coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
  const        0.174428     0.172521      1.011    0.3191
  xrate_1     -0.0582389    0.0612668    -0.9506   0.7725
  d_xrate_1    0.400819     0.167458      2.394    0.0224   **

  AIC: -41.5297   BIC: -36.697   HQC: -39.826

*********************************

   with constant and trend
```

```
   model: (1-L)y = b0 + b1*t + (a-1)*y(-1) + ... + e
   1st-order autocorrelation coeff. for e: -0.011
   estimated value of (a - 1): -0.0605989
   test statistic: tau_ct(1) = -1.00215
   asymptotic p-value 0.9422

Augmented Dickey-Fuller regression
OLS, using observations 1991:3-2000:3 (T = 37)
Dependent variable: d_xrate

              coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
  const        0.121026     0.174493      0.6936   0.4928
  xrate_1     -0.0605989    0.0604690    -1.002    0.9422
  d_xrate_1    0.348321     0.169477      2.055    0.0478   **
  time         0.00288669   0.00207752    1.389    0.1740

  AIC: -41.6335   BIC: -35.1898   HQC: -39.3618
```

As expected, in the second model with constant and trend, time is insignificant therefore we assume that `xrate` is described by the model $\Delta xrate_t = \mu_0 + \rho\, xrate_{t-1} + \gamma_1 \Delta xrate_{t-1} + \varepsilon_t$ with insignificant constant. We test $H_0 : \rho = 0$ and, since $p-$value equals 0.7725 (which implies that `xrate` has a unit root), we conclude that `xrate` is an ARIMA(1,1,0) process. To forecast, go to Modell Time seriesl ARIMA...l choose `xrate` as Dependent variable, AR order 1, Difference 1, MA order 0, no constantl OK (see Fig. 4.22, right). The two forecasts there are quite different!

**Exercise 4.4**. Consider the quarterly U.S. real seasonally adjusted `gnp` (gross national product) from 1947:01 to 2002:03 contained in the second column of gnp47.txt (the first column contains `quart` (=1947.00, 1947.25, 1947.50,…, 2002.50).

1. Create two new series, $l\_gnp_t = \log(gnp_t)$ and log differences of gnp[11], $ld\_gnp_t = \log(gnp_t) - \log(gnp_{t-1})$. Plot `l_gnp` and `ld_gnp`. Which of the two series seems to be stationary?

2. Create a new series `time` (use Addl Time trend) (=1, 2, 3, ...). Estimate two OLS models, $l\_gnp = \alpha + \beta_1 time + \varepsilon$ and $l\_gnp = \alpha + \beta_1 quart + \varepsilon$. Explain which $\beta_1$ gives an average percentage <u>quarterly</u> growth of `gnp` and which one the <u>annual</u> growth rate[12]. How do these two numbers compare to the mean value of `ld_gnp`?

3. Plot the sample ACF and PACF of the quarterly growth rate `ld_gnp`. Inspecting the sample ACF and PACF, you might feel that the ACF is cutting off at lag 2 and the PACF is tailing off. This would suggest that the gnp growth rate `ld_gnp` follows an MA(2) process, or `l_gnp` follows an ARIMA(0, 1, 2) model. Another variant to explain the correlogram is to suggest that its ACF is tailing off and the PACF is cutting off at lag 1. This suggests an AR(1) model for the growth rate or ARIMA(1,1,0) for `l_gnp`. Estimate both ARIMA models (with constants)[13].

[11] The economic meaning of `ld_gnp` is a quarterly percentage growth of `gnp`.

[12] Recall that $\beta_1$ in, say, the equation $l\_gnp = \alpha + \beta_1 quart + \varepsilon$ means the percentage growth of $gnp$ when $quart$ increases by 1.

[13] Both models are nearly the same – in models window go to Graphsl Fitted, actual plotl Against time.

4. Forecast l_gnp 12 quarters ahead using both ARIMA models, compare the forecasts. Forecast the OLS model $l\_gnp = \alpha + \beta_1 time + \varepsilon$ 12 quarters ahead (surprisingly, this forecast does not differ much from the previous ones). ◄◄

This ends our **very short** introduction to statistics. Still a long way to go ...

## References


| | |
|---|---|
| [A] | Adkins, Lee C. Using gretl for Principles of Econometrics, 4th ed.<br>http://www.learneconometrics.com/gretl/index.html |
| [gretl] | http://gretl.sourceforge.net/ |